STOCHASTIC SIR HOUSEHOLD EPIDEMIC MODEL WITH MISCLASSIFICATION.

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

UNIVERSITY OF KENT

By

UMAR MALLAM ABDULKARIM

NOVEMBER 2016

Acknowledgements.

I wish to express my profound gratitude to my supervisors, Professor Martin Ridout and Dr Owen D. Lyne and the entire staff of the School of Mathematics, Statistics and Actuarial science for their encouragement and guidance in seeing that this work is sucessfully completed. I will ever remain grateful for their support including those of my colleagues, students without which this achievement might not have been possible. Thank you all for the useful time we had together.

My gratitude also goes to my wife, Hajiya Khadiza (Nadi) Abdulkarim and the children for their patience and understanding during which my absence from home was least desired. Your sacrifices and those of the children have been of tremendous benefit to the family.

I will like to thank my sponsors, the federal government of Nigeria, management of the board of TETFUND and the management of the Nasarawa state University Keffi for the opportunity and funding to undergo this programme.

Abstract.

Often data from infectious disease are subject to classification errors, such as susceptible individuals classified as infectives or vice versa. These kinds of classification error may lead to imprecise record of the number of individuals infected in each household and therefore unreliable results of inference from such data. It then becomes necessary to adjust our parameter estimation methods to cope with such errors and obtain precise maximum likelihood estimates that reflect the true parameter values and model that best fit the final size epidemic data.

In this work, we have proposed a theoretical framework leading to misclassification error probabilities from the SIR household epidemic and procedures on how the inference should be handled in the face of these errors, given the following scenarios,

(i) When there is no misclassification error in the data, (misclassification probability=0), in which case, the true positives are classified as such, while true negatives are also correctly classified as such.

(ii) When the false negative and positive misclassification probabilities are the same.

(iii) When these misclassification probabilities are different from each other.

Using maximum likelihood inference, we simulated household final size epidemic data with error and showed that the parameters from the models with misclassification error in (ii) and (iii) including the error rate can be correctly estimated just as in the case without error.

Since misspecification may wrongly be taken for miscclassification of the epidemic data, we examined the effects of misspecification of the infectious period distribution on the estimates, considering the three scenarios listed above to see how the behaviours of the estimates differ from those of misclassification of the epidemic data.

The Pearson chi-square goodness of fit test and the Kolmogorov-Smirnov goodness of fit test are employed to assess the goodness of fit of the models given three scenarios in (i)-(iii) referred to as the two, three and four dimensional models respectively in relationship to the number of parameters in the models. The three models are found to sufficiently fit the two dimensional final size epidemic data. The three and four dimensional models perform well on the three dimensional final size epidemic data, while the two dimensional model failed to sufficiently fit the three dimensional final size epidemic data when the misclassification probability is not close to 0.

Similar behaviours from the two dimensional model are observed on the four dimensional final size epidemic data, while the three dimensional model performs well on the four dimensional final size epidemic data when the misclassification probabilities are close to each other. The four dimensional model performs well on the four dimensional final size data for any choice of the misclassification probabilities in the permissible region, $[0, 0.5)$.

These behaviours are further examined from the mean and variance of the Pearson chi-square goodness of fit statistics of the three scenarios (i)-(iii) and those of the proportion of the simulations rejected at 5% level of significance from the Pearson chi-square goodness of fit test. We see that with increasing misclassification probabilities in the permissible region, $[0, 0.5)$, the proportion of the simulations rejected for the two and three dimensional models tend to 1 respectively, while those of the four dimensional model remains consistently stable around 5% as theoretically expected.

Also, we employed the chi-square difference goodness of fit test and the Kolmogorov-Smirnov goodness of fit test, given the three scenarios and found the behaviours of the models to be consistent with our earlier studies.

We employed these procedures to the [1] Tecumseh Michigan influenza A(H3,N2) epidemic data and [28] Seattle Influenza $1975-1976$ B(H1N1), $1978-1979$ A(H1N1) epidemic data and found that the three models sufficiently fit the final size epidemic data.

We have shown that the four dimensional model outperforms the two and three dimensional models and that the two and three dimensional model are only useful if the misclassification probabilities are close to 0 as in the case of the two dimensional model and when they

are close to each as in the case of the three dimensional model respectively.

However if the misclassification probabilities are far apart then the two and three dimensional models struggle fitting to the four dimensional final size epidemic data. Hence the need for the four dimensional model.

Also we see that in the presence of misspecification of the models, the two dimensional model is better than the complex models if the epidemic data are not misclassified otherwise the complex models are better.

Contents

1 Introduction. 1

1.1 Overview. . . . 1

1.2 Motivation of the study. . . . 4

1.3 Introduction to the thesis. . . . 6

1.4 Background of study. . . . 8

1.4.1 Empirical Approach to the study of Infectious Disease. . . . 8

1.4.2 Work on stochastic epidemic modelling. . . . 9

1.5 Misclassification of household epidemic data. . . . 11

1.5.1 Epidemic modelling in the presence of misclassification. . . . 13

1.5.2 Literature on modelling misclassified finite count data. . . . 14

1.6 The stochastic SIR epidemic model. . . . 16

1.7 Branching process. . . . 17

1.8 Convergence of the general stochastic epidemic. . . . 19

1.9 Final size household epidemic data. . . . 20

1.10 Dimensionality of household epidemic data. . . . 21

1.11 The Gontcharoff polynomial. . . . 22

2 The stochastic SIR household epidemic model. 24

2.1 Introduction. . . . 24

2.2 Household structure. . . . 25

2.3 Two level mixing epidemic model. . . . 25

2.4 Branching process for two level mixing epidemic model. . . . 27

2.5 Community based SIR household epidemic model with temporary immunity. . 28

2.6 Community based SIR household epidemic model with permanent immunity. 28

2.6.1 Calculation of the final size probabilities. 29
2.7 Threshold parameter. 30
2.8 Mean final size of single household epidemic. 31
2.9 Numerical simulations. 33
2.10 Inference on the parameters. 35
2.11 Global epidemic. 35
2.12 Maximum likelihood estimation. 37

3 Theoretical properties of the parameters of the stochastic SIR household epidemic model. 39

3.1 Introduction. 39
3.2 The mean final size of single household epidemic. 39
3.3 Properties of β_k for small and large local infection rates. 41
3.4 The mean final size of the single household epidemic for small λ_L. 43
3.5 The mean final size of the single household epidemic, for large local infection rates. 44
3.6 Further properties of the mean final size. 45
3.7 Properties of the threshold parameter for small and large local infection rates. 50
3.8 Proportion of the initial susceptibles that are ultimately infected. 52
3.9 Proportion of the initial susceptibles that are ultimately infected at the lower boundary of the local infection rate. 55
3.10 Proportion of the initial susceptibles that are ultimately infected near the upper boundary of the local infection rate. 56
3.11 Theoretical properties of the Gamma function and global epidemic. 57

4 Fitting the SIR household model to final size epidemic data. 60

4.1 Introduction. 60
4.2 Model fitting to the two dimensional final size data. 60
4.3 Replication of published results. 63
4.4 Simulation and inference. 64

4.5 Plots of the estimates with minimum epidemic size of 10. 66
4.5.1 Table of parameter estimates and other statistics when the minimum epidemic size is 10. 67
4.6 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 50. 68
4.6.1 Table of parameter estimates and other statistics when the minimum epidemic size is 50. 69
4.7 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 100. 70
4.7.1 Table of parameter estimates and other statistics when the minimum epidemic size is 100. 71
4.8 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 1000. 71
4.9 Parameter estimates with minimum epidemic size of 1000. 73
4.9.1 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.0446$ and $\lambda_G = 0.1955$ with minimum epidemic size of 1000. 74
4.9.2 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.13$ and $\lambda_G = 0.17$ with minimum epidemic size of 1000. . . 76
4.9.3 Plots of the estimates of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.1$ and $\lambda_G = 0.29$ with minimum epidemic size of 1000. . . . 77
4.9.4 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.25$ and $\lambda_G = 0.39$ with minimum epidemic size of 1000. . . 78

5 Stochastic SIR household model for misclassified data. 80
5.1 Introduction 80
5.2 The SIR household epidemic model with two different misclassification probabilities. 81
5.3 The three dimensional final size epidemic model. 88
5.3.1 Maximum likelihood estimation. 89

5.4 Numerical simulations and inferences on the three and four dimensional final size epidemic data. 90
5.4.1 Fitting the three models to data from the four dimensional model. . . 90
5.4.2 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$. 93
5.4.3 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$. 94
5.4.4 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2$. 95
5.5 Numerical simulations and inferences. 97
5.6 Comparison of the models. 98
5.6.1 Simulations with the theoretical parameter, $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.4316$. 99
5.6.2 Simulations with theoretical parameters, $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$. 101
5.7 Summary of behaviour of the models. 103
5.8 Simulations and inferences of the three models. 104
5.8.1 Fitting the two, three and four dimensional models to the three dimensional final size epidemic data. 106
5.8.2 Fitting the two, three and four dimensional models to the three dimensional simulated final size epidemic data, when $\varepsilon = 0.01$. 106
5.8.3 Fitting the two, three and four dimensional models to the three dimensional simulated final size epidemic data, when $\varepsilon = 0.02$. 107
5.8.4 Fitting the two, three and four dimensional models to three dimensional simulated final size epidemic data, when $\varepsilon = 0.2$. 107
5.9 Table of mean, standard deviation and root mean square error of the estimates for the two, three and four dimensional models, when $\varepsilon = 0.01$, 0.02 and $\varepsilon = 0.2$. 109
5.10 Simulations and inferences of the two and three dimensional models for $z \in [0, 1]$. 110

5.10.1 Plots of the RMSE of the Parameter estimates when, $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\pi = 0.8999$, $z = 0.2144$, $R_* = 1.1653$. 112
5.10.2 Plots of the RMSE of the parameter estimates when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$. 113
5.11 Summary of performance of the two, three and four dimensional models on final size epidemic data. 114

6 Chi-square goodness of fit test. 116
6.1 Introduction. 116
6.2 Computation method of the Pearson chi-square goodness of fit statistic. . . . 118
6.3 Degrees of freedom of the Pearson chi-square goodness of fit test. 120
6.4 Likelihood ratio chi-squared goodness of fit test. 120
6.5 Kolmogorov-Smirnov test. 121
6.6 Proportion of the simulations rejected from the Pearson chi-square goodness of fit test. 122
6.7 Pearson chi-square goodness of fit test on two dimensional final size epidemic data. 124
6.8 Numerical simulations on two dimensional final size epidemic data. 126
6.8.1 The Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests on two dimensional final size epidemic data. 126
6.8.2 Table of mean and variance of the Pearson chi-square test on the two dimensional final size epidemic data. 127
6.9 The Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests on the three dimensional final size epidemic data. 129
6.9.1 When the misclassification probability $\varepsilon = 0.1$. 130
6.9.2 When the misclassification probability $\varepsilon = 0.3$. 131
6.9.3 Table of mean and variance of the Pearson chi-square goodness of fit statistic on the three dimensional final size epidemic data. 132

6.10 Plots of the mean and variance of the Pearson chi-square goodness of fit statistic on the three dimensional final size data. 133
6.11 The Pearson chi-square goodness of fit tests on the four dimensional final size epidemic data. 136
6.11.1 When the misclassification probabilities are $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$. . 137
6.11.2 When the misclassification probabilities are $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$. . 138
6.11.3 When the misclassification probabilities are $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$ 140
6.11.4 When the misclassification probabilities are $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$ 142
6.11.5 When the misclassification probabilities are $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$ 144
6.11.6 When the misclassification probabilities are $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$ 146
6.12 Table of mean and variance of the Pearson chi-square goodness of fit statistics of the three models on the four dimensional final size epidemic data. 148
6.13 Plots of the mean and variance of the Pearson chi-square goodness of fit statistic. 150
6.13.1 Exploring the estimates along the diagonals, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$, theoretical parameters corresponding to $z = 0.7298$, 0.2144 respectively. 150
6.13.2 Exploring the estimates along the vertical axis of the misclassification probability region. 153
6.13.3 Exploring the estimates along the horizontal axis of the misclassification probability region. 155
6.14 Fitting the three models to [1] Tecumseh Michigan Influenza A(H3N2) epidemic data. 156
6.15 Analyses of the Seattle influenza datasets. 157
6.15.1 Analyses of the epidemic datasets. 158
6.16 Fitting the three models to the Seattle household epidemic data. 159
6.16.1 The 1975-1976 Seattle B(H1N1) influenza epidemic. 159
6.16.2 The 1978-1979 Seattle A(H1N1) influenza epidemic. 160
6.17 Discussion and Comments. 161

7 Hypothesis test between the models. 164

7.1 Introduction. . . . 164

7.2 Chi-square difference test. . . . 165

7.3 Kolmogorov-Smirnov test. . . . 166

7.4 Proportion of the simulations rejected from the chi-square difference test. . . . 166

7.5 Chi-square difference and the Kolmogorov-Smirnov tests on the two dimensional final size epidemic data. . . . 167

7.6 Table of mean and variance of the chi-square difference tests on the two dimensional final size epidemic Data. . . . 169

7.7 Chi-square difference and the Kolmogorov-Smirnov tests on the three dimensional final size epidemic data. . . . 170

7.8 Table of mean and variance of the chi-square difference statistic on the three dimensional final size epidemic Data. . . . 174

7.9 Plots of the mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data. . . . 175

7.10 The chi-square difference and Kolmogorov-Smirnov tests on the four dimensional final size epidemic data. . . . 178

7.11 Table of mean and variance of the chi-square difference statistic. . . . 190

7.12 Plots of the mean and variance of the chi-square difference statistic on the four dimensional final size epidemic data. . . . 191

7.13 Fitting the three models to [1] Tecumseh Michigan Influenza A(H3N2) epidemic data using chi-square difference statistic. . . . 194

7.14 Fitting the three models to [28] Seattle Influenza epidemic data using chi-square difference statistic. . . . 195

7.15 Discussion 196

8 Estimation in the presence of model misspecification. 197

8.1 Introduction 197

8.2 Simulating epidemic data with exp(4.1) and estimating model parameters with Gamma$(2, 4.1/2)$ infectious period distributions. 198
8.3 Simulating epidemic data with Gamma$(2, 4.1/2)$ and estimating model parameters with exp(4.1) infectious period distributions. 200
8.4 Discussion and comments. 201
8.5 Effects of misspecification on the estimates of the three models from two dimensional epidemic data. 201
8.6 When the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution. 204
8.7 Discussion and comments. 207
8.8 Misspecification in the face of misclassification. 207
8.9 When the epidemic data is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions. 208
8.10 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions. 212
8.11 Conclusion and comments. 216
8.12 Misspecification in the face of different misclassification Probabilities. 216
8.13 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions. 217
8.13.1 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions. 220
8.14 Conclusion and comments. 224

9 Summary, Conclusion and Extensions. 226

9.1 Introduction. 226

9.2 Summary of Work. 226
9.3 Discussion. 231
9.4 Possible Extension. 235
9.5 Overall Conclusion. 236
9.6 Limitation of the Study. 238
9.7 Recommendation. 238

Bibliography **240**

List of Figures

2.1 Histogram of 1000 simulations of household epidemic with Gamma$(2, 2.05)$ infectious period distribution and parameter estimates from [1] but fifty times its population size. 34

3.1 The beta function with increasing λ_L. 43
3.2 The mean fnal size as function of the local infection rate. 46
3.3 The mean final size as function of the number of initial infectives 48
3.4 The mean final size as function of number of the initial susceptibles 49
3.5 The threshold parameter with varying local infection rate. 52
3.6 The proportion of the initial susceptible ultimately infected at the end of the epidemic in the presence of varying values of π. 54

4.1 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and minimum epidemic size of 10. 66
4.2 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and Minimum Epidemic size of 50. 68
4.3 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and Minimum Epidemic size of 100. 70
4.4 Histograms of number infected from simulations of household epidemic, with population sizes of 1414, and 70700 respectively, minimum epidemic size of 1 and simulation runs of 1000. 73

4.5 Plots of the Estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.0446$, $\lambda_G = 0.1955$ and minimum epidemic size of 1000. 75
4.6 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.13, \lambda_G = 0.17$ and minimum epidemic size of 1000. 76
4.7 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.1, \lambda_G = 0.29$ and minimum epidemic size of 1000. 77
4.8 Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.25, \lambda_G = 0.39$ and minimum epidemic size of 1000. 78

5.1 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$ 93
5.2 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$ 94
5.3 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2$ 95
5.4 Plots of the root mean square error of the maximum likelihood estimates of the three models when, $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.4316$. 101
5.5 Plots of the root mean square error of the maximum likelihood estimates for the three models, when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$. 103
5.6 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.01$ 106
5.7 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.02$. 107
5.8 Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.2$. 108
5.9 Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $z = 0.2144$. 112

5.10 Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$. 113

6.1 Density histograms of the Pearson chi-square goodness of fit and the Kolmogorov-Smirnov goodness of fit tests on the models two dimensional final size epidemic data. 127

6.2 Density histograms of the Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests of the three models on three dimensional final size epidemic data, when $\varepsilon = 0.1$. 130

6.3 Density histograms of the Pearson chi-square and the Kolmogorov Smirnov goodness of fit tests of the three models on the three dimensional final size epidemic data when $\varepsilon = 0.3$. 131

6.4 Plots of the mean and variance of the Pearson chi-square goodness of fit statistics for the three models when $\lambda_L = 0.1$, $\lambda_G = 0.29$ and $\lambda_L = 0.2$, $\lambda_G = 0.12$. 134

6.5 Plots of the proportion of the simulations rejected at 5% significance from the Pearson chi-square goodness of fit tests. 135

6.6 Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution function of the Pearson chi-square goodness of fit statistic with their hypothesized distributions for the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$.. 137

6.7 Density histogram of the Pearson chi-square, the likelihood ratio chi-squared goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$. 139

6.8 Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plot of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$. 141

6.9 Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$. 143

6.10 Density histograms of Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$. 145

6.11 Density histograms of chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$.. 147

6.12 Plots of the mean and variance of the chi-square goodness of fit statistics for the three models when the estimates are explored along the diagonals of the misclassification region for theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$. 151

6.13 Plots of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test to four dimensional final size epidemic data for theoretical parameters corresponding to $z = 0.2144$. and $z = 0.7298$ respectively. 152

6.14 Plots of the mean and variance of the chi-square goodness of fit statistics for the three models with $\varepsilon_{FP} = 0.01$, while varying ε_{FN} with step size of 0.01, for theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$. 154
6.15 Plots of the mean and variance of the chi-square goodness of fit statistics for the three models with $\varepsilon_{FN} = 0.01$, while varying ε_{FP} with step size of 0.01, for theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$. 156

7.1 Density histograms of the chi-square difference statistic on two dimensional final size epidemic data and plots of the empirical distribution of the chi-square difference statistic. 168
7.2 Density histograms of the chi-square difference statistic on the three dimensional final size epidemic data and those of the empirical and cumulative distribution functions when $\varepsilon = 0.1$. 171
7.3 Density histograms of the chi-square difference statistic on the three dimensional final size epidemic data and those of the empirical and cumulative distribution functions when $\varepsilon = 0.3$. 173
7.4 The mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data. 176
7.5 Proportion of the simulations rejected at 5% significance from the chi-square difference test for $z = 0.7298$ and $z = 0.2144$ when it is the three dimensional final size epidemic data. 177
7.6 Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$. 179
7.7 Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$ 181

7.8 Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions, when $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$ 183

7.9 Density histogram of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$ 185

7.10 Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of their empirical distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$ 187

7.11 Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of their empirical distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$ 189

7.12 Plots of the mean and variance of the chi-square difference statistic for the three models in the parameter estimates are explored along the diagonal, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$ over the misclassification region with theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$. respectively. 192

7.13 Proportion of the simulations rejected at 5% significance from the chi-square difference test with theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$, when the true data is the four dimensional final size epidemic data. 193

8.1 Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution and when the epidemic data is simulated with exp(1.4) infectious period distribution. 199

8.2 Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution. 200

8.3 Plots of the estimates from the three models when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, the parameters estimated with Gamma$(2, 4.1/2)$ infectious period distribution. 202

8.4 Plots of the estimates from the three models when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, the parameters estimated with exp(4.1) infectious period distribution. 205

8.5 Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.01$. 208

8.6 Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.02$. 209

8.7 Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.2$. 210

8.8 Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.01$. 212

8.9 Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.02$. 213

8.10 Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.2$. 214

8.11 Plots of the estimates using Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.02$, $\varepsilon_{FF} = 0.1$, ... 217

8.12 Plots of the estimates using Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.3$, $\varepsilon_{FF} = 0.2$, ... 218
8.13 Plots of the estimates using Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.2$, $\varepsilon_{FF} = 0.2$, ... 219
8.14 Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.02$, $\varepsilon_{FF} = 0.1$, ... 221
8.15 Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.3$, $\varepsilon_{FF} = 0.2$, ... 222
8.16 Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.2$, $\varepsilon_{FF} = 0.2$, ... 223

List of Tables

1.1 Household epidemic data 21
1.2 Tecumseh Michigan Influenza A(H3N2) Epidemic Data 21

4.1 Table of Comparison of Parameter Estimates 64
4.2 Pairs of the local and global infection rates with their corresponding theoretical parameters 65
4.3 Mean of the parameter estimates for theoretical parameters corresponding to $z = 0.1775$ and minimum epidemic size of 10. 67
4.4 Mean of the parameter estimates for theoretical parameters corresponding to $z = 0.1775$ with minimum epidemic size of 50. 69
4.5 Mean of the parameter estimates for theoretical paramters corresponding to $z = 0.1775$ with minimum epidemic size of 100. 71
4.6 Table of comparison of the mean, standard deviation and mean square error of the estimates using the minimum epidemic size of 1000 and simulation runs of 1000. 73
4.7 Mean of the parameter estimates from the two dimensional model and theoretical parameters in table 4.2. 79
4.8 Standard deviation of the parameter estimates from the two dimensional model with theoretical parameters in table 4.2. 79
4.9 Root mean square error of the parameter estimates from the two dimensional model with theoretical parameters in table 4.2.. 79

5.1 Table of the mean of the parameter estimates of the three models. 96
5.2 Table of the standard deviation of the parameter estimates of the three models. 96

5.3 Table of the root mean square error of the parameter estimates of the three models. 96
5.4 Mean of the parameter estimates of the two, three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model. 109
5.5 Standard deviation of the parameter estimates of the two and three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model. 110
5.6 Root mean square error of the parameter estimates of the two and three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model. 110
5.7 Table of comparison of optimisations and models on the two, three and four dimensional simulated final size epidemic data. 115

6.1 Table of the expected number of i infected in household of the given sizes . . 118
6.2 Table of mean and standard deviation of the Pearson chi-square of fit statistic of the models on two dimensional final size epidemic data 127
6.3 Table of the proportion of the simulations rejected from the two dimensional final size epidemic data. 128
6.4 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the two dimensional final size epidemic data. 128
6.5 Summary of the Kolmogorov-Smirnov test for the upper 5% points for the three dimensional final size epidemic data when $\varepsilon = 0.1$. 131
6.6 Summary of the Kolmogorov-Smirnov test for the upper 5% points for the three dimensional final size epidemic data when $\varepsilon = 0.3$. 132
6.7 Table of the mean and variance of the Pearson chi-square goodness of fit statistic on the four dimensional final size epidemic data 132
6.8 Table of the proportion of the simulations rejected from the Pearson chi-square test for the three dimensional final size epidemic data. 133

6.9 Table of misclassification probabilities 1 to 6. 136
6.10 Summary of the Kolmogorov-Smirnov goodness of fit test with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0.2$. 138
6.11 Summary of the Kolmogorov-Smirnov test goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0$. 140
6.12 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.01$, $\varepsilon_{FP} = 0.02$ 142
6.13 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$. 144
6.14 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.3$. 146
6.15 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$. 148
6.16 Table of mean and variance of the Pearson chi-square goodness of fit statistics on the four dimensional final size epidemic data 149
6.17 Table of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test on the four dimensional final size epidemic data. 149
6.18 Table of the parameter estimates of the models from [1] Final size data 157
6.19 Influenza B(H1N1) 1975-1976 final size data. 158
6.20 Influenza A(H1N1)1978-1979 final size data. 158
6.21 Estimates from the 1975-1976 Seattle B(H1N1) influenza epidemic 158
6.22 Estimates from the 1975-1976 Seattle B(H1N1) influenza epidemic 159

6.23 Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(1, 4.1)$ infectious period distribution from the 1975-1976 B(H1N1) influenza epidemic. 159
6.24 Parameter estimates and Pearson chi-square and likelihood ratio chi-squared goodness of fit statistics with Gamma$(2, 4.1/2)$ infectious period distribution from the $1975 - 1976$ B(H1N1) influenza epidemic. 160
6.25 Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(5, 4.1/5)$ infectious period distribution from 1975-1976 B(H1N1) influenza epidemic. 160
6.26 Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(1, 4.1)$ infectious period distribution from 1978-1979 A(H1N1) influenza epidemic. 160
6.27 Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(2, 4.1/2)$ infectious period distribution from 1978-1979 A(H1N1) influenza epidemic. 161
6.28 Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(5, 4.1/5)$ infectious period distribution from 1978-1979 A(H1N1) influenza epidemic. 161

7.1 Table of mean and variance of the chi-square difference tests on the two dimensional final size epidemic data. 169
7.2 Proportion of the simulations rejected from the chi-square difference test at 5% significance from the two dimensional epidemic data 169
7.3 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the two dimensional final size epidemic data. 170
7.4 Table of summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the three dimensional final size epidemic data when $\varepsilon = 0.1$. 172
7.5 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the three dimensional final size epidemic data when$\varepsilon = 0.3$. . . . 174

7.6 Proportion of the simulations rejected rom the chi-square difference test at 5% significance from the three dimensional final size epidemic data. 174
7.7 The mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data with misclassification probabilities, ε = 0.0, 0.1, 0.3.. 175
7.8 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$, ε_{FP} = 0.2 180
7.9 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when ε_{FN} = 0.2, $\varepsilon_{FP} = 0$ 182
7.10 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when ε_{FN} = 0.01, $\varepsilon_{FP} = 0.02$ 184
7.11 Summary from the Kolmogorov-Smirnov test for the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$ 186
7.12 Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when ε_{FN} = 0.2, $\varepsilon_{FP} = 0.3$ 188
7.13 Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when ε_{FN} = 0.3, $\varepsilon_{FP} = 0.2$ 190
7.14 The mean and variance of chi-square difference statistic on the four dimensional final size epidemic data simulated with misclasification probabilities in table 6.9.190
7.15 Table of chi-square difference statistic for the three models and their corresponding P-values. 194
7.16 Table of chi-square difference statistic and their corresponding P-values for the three models from the Seattle $1975 - 1976$ B(H1N1) Influenza epidemic with $T_I = \text{Gamma}(a, b)$ infectious period distribution. 195

7.17 Table of chi-square difference statistic and their corresponding P-values for the three models from the Seattle $1978-1979$ A(H1N1) Influenza epidemic with $T_I = \text{Gamma}(a, b)$ infectious period distribution. 195

8.1 Table of mean, standard deviation and root mean square error of the estimates when the epidemic data is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions. 199

8.2 Table of mean, standard deviation and root mean square error of the estimates when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions. 201

8.3 Table of mean of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution and estimated with Gamma$(2, 4.1/2)$ infectious period distribution. 203

8.4 Table of standard deviation of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, estimated with Gamma$(2, 4.1/2)$ infectious period distribution. 203

8.5 Table of the root mean square error of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, estimated with Gamma$(2, 4.1/2)$ infectious period distribution. 204

8.6 Table of mean of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution. . 206

8.7 Table of standard deviation of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution. 206

8.8 Table of root mean square error of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution. 207

8.9 Mean of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution. 211

8.10 Standard deviation of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution. 211

8.11 Root mean square error of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution. 211

8.12 Mean of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is with simulated with Gamma$(2, 4.1/2)$ infectious period distribution. 215

8.13 Standard deviation of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution. 215

8.14 Root mean square error of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution. 216

8.15 Table of mean of the parameter estimates when the epidemic is simulated with exp(4.1) and estimated and Gamma$(2, 4.1/2)$ infectious period distributions. 220

8.16 Table of the standard deviation of the parameter estimates when the epidemic is simulated with exp(4.1) and estimated and Gamma$(2, 4.1/2)$ infectious period distributions. 220

8.17 Table of the root mean square error of the parameter estimates when the epidemic is simulated with exp(4.1) and estimated and Gamma(2, 4.1/2) infectious period distributions. 221
8.18 Table of mean of the parameter estimates when the epidemic is simulated with Gamma(2, 4.1/2) and estimated and exp(4.1) infectious period distributions. 224
8.19 Table of the standard deviation of the parameter estimates when the epidemic is simulated with Gamma(2, 4.1/2) and estimated and exp(4.1) infectious period distributions. 224
8.20 Table of the root mean square error of the parameter estimates when the epidemic is simulated with Gamma(2, 4.1/2) and estimated and exp(4.1) infectious period distributions. 224

Chapter 1

Introduction.

1.1 Overview.

This work is concerned with the study of stochastic models of infectious diseases in a closed population partitioned into small groups. These may represent people living together in the same dwellings and we will call them households. Each household is made up of susceptible, infective and removed individuals. A susceptible individual is one who can be infected with the disease, an infective is one who has the disease and a removed individual is one who has been removed (because it has recovered and is immune from further re-infection of the disease under discussion or isolated or has died).

Stochastic models are the natural tools for studying infectious diseases, as they can incorporate randomness in the transmission pattern of infectious diseases, especially in small populations. Their usage in modelling infectious diseases has a long history, which is outlined briefly in section 1.4.

There are many of these models available with applications to transmission of infectious disease in human and animals. Among them is the deterministic SIR epidemic model of [44], where the acronym SIR stands for susceptibles, infectives and removed individuals respectively.

The deterministic SIR epidemic model of [44] assumes homogeneous mixing between individuals in a constant population (no birth/death or migration/immigration) [27, 30]. An

individual contacted is immediately infectious for a period T_I referred to as the infectious period, after which it recovers and becomes immune or dies from the infection. The SIR epidemic model with exponentially distributed infectious period was first studied by [20]. Various extensions and generalisations have been proposed by other research workers. One of which, that of [9], forms the basis of this research.

Discussions of the general stochastic epidemic model, its theoretical properties and its extension by [9] is provided in sections 1.6 and 2.1 respectively.

Most infectious disease data are subject to error during their collection. This may be caused by incorrect classification of individuals' health state and hence lead to unreliable estimates of the parameters and inadequate fit of the model to data. It then becomes necessary to adjust our inferences to cope with this circumstance by providing suitable methods that take account of these errors and still give precise estimates of the parameters and hence a more reliable model fit that mimics our data.

Our focus in this work is fitting data from epidemics of infectious diseases to the stochastic SIR household epidemic model taking into consideration cases when the epidemic data is subject to classification error.

We will be estimating the model parameters using the likelihood approach and maximum likelihood inference in [1]. This likelihood function will be referred to as the approximate likelihood in our model, as the assumption on which it is based is not consistent with that of [9]. The precision of the maximum likelihood estimators is assessed from their mean, standard deviation and root mean square error. Plots of the root mean square error of the parameter estimates for a range of percentage misclassification errors are studied to give insights into the behaviours and properties of the model.

We will evaluate the performance of two, three and four dimensional models for small and large percentage error (misclassification probabilities) in the permissible region $0 \leq \varepsilon < 0.5$ and then assess the fitted models to the final size epidemic data using goodness of fit statistics, by comparing the observed and the expected number infected for discrepancies or otherwise.

In order to achieve this, we have first developed a theoretical framework leading to misclassification error in the household epidemic data for the two cases in which it can occur

namely,

(a) When the misclassification probabilities are the same. That is, the probability of making false negative classification error is the same as that of making false positive classification error. A false negative classification error occurs when a true positive is observed to be negative, while a false positive occurs when a true negative is observed to be positive. The theoretical basis leading to this kind of misclassification probability is discussed in section 5.3 of chapter 5.

(b) When the misclassification probabilities errors are different from each other. The theoretical framework leading to these misclassification probabilities in the final size epidemic data is developed and discussed in sections 5.2 and 5.5 of chapter 5.

From (a) and (b) we see that the models are nested as follows,

(i) If every infective individual is correctly observed as infective and every susceptible is observed correctly as susceptible then the probability of making these classification errors is simply zero. When this happens there will be no error (noise) in the household epidemic data and the likelihood function will only be a function of two parameters, the local infection rate and the probability of avoiding infection from outside the household. This is further discussed and explored in section 4.2 of chapter 4.

(ii) If infectives are wrongly classified as susceptible and susceptibles wrongly classified as infectives, such that the probability of making these classification errors are the same; then case (a) is realised.

(iii) If infectives are wrongly classified as susceptibles and susceptibles as infectives with different probabilities of making these classification errors; then case (b) is realised.

Using simulation studies and the appropriate numerical optimization schemes, we explored the parameter estimates of these models and examined their precision by computing their mean, standard deviation, mean square error and root mean square error. These can be found in chapters 4 and 5 respectively.

In chapters 6 and 7, we employed goodness of fit statistics to test for fitness of the model to the final size epidemic data also used them to analyse the [1] and [28] epidemic data.

In chapter 8, we studied the effects of misspecification of the infectious period distribution

on the estimates of the stochastic SIR household epidemic model in the face of no misclassification and misclassification of the epidemic data.

1.2 Motivation of the study.

Deaths from microorganism-induced epidemics are often in the range of thousands of people and therefore a threat to the continuous existence of humanity [3]. Sometimes, this large number of deaths may be attributable to inadequate treatment regimes, intervention strategies, low level of literacy and poverty especially in the developing world, to stop the epidemic from spreading when it is started. For example, Plague, otherwise called black death, is known to have been responsible for a widespread pandemic with high mortality during the fourteenth century [3].

Europe suffered an estimated 100 million deaths from the so called black death alone [17]. The Aztecs lost half of their population to a smallpox epidemic in 1520 leading to the downfall of its empire, while Russia suffered from an epidemic of typhus between 1918 and 1921, with a death rate of about 25% of its population [17]. The 1919 world pandemic of influenza killed over 20 million people in 12 months alone [17].

Cholera is an acute infection that spreads rapidly where living conditions are crowded, water sources are unprotected and there is lack of safe disposal of faeces [3]. These are conditions commonly faced by people living in poor countries of the world and also in refugee camps. For example, in a refugee camp in the Democratic Republic of Congo, an estimated $58,000 - 80,000$ cases were recorded within one month in 1994 with $23,800$ deaths [3].

In recent times, the epidemics of HIV/AIDS have been the focus of the World Health Organisation to bring the transmission of the disease in countries with high levels of prevalence under control, especially the developing countries where public heath care systems are inadequate to cope with large numbers of infections of the disease. For example UNAIDS and World Bank reports indicate that the HIV/ AIDS epidemic was responsible for 8.6% of death from infectious diseases in the developing world and that in the year 2020 it will be responsible for 37.1% of such deaths among adults between the ages of 15 and 59 years [51].

Just as the world is still grappling with the epidemic of HIV/AIDS, an Ebola epidemic emerged, ravaging the West African subregion. Countries like Guinea, Sierra Leone and Liberia were most affected with high numbers of cases suspected, probable and confirmed, including deaths. For example, according to the World Health Organisation situation report of 28th April 2016 on Ebola virus diseases in the three countries, $29,616$ suspected, probable and confirmed cases were reported, with $15,227$ laboratory confirmed cases and $11,310$ deaths, while in other affected countries, 36 suspected, probable and confirmed cases were reported with 34 laboratory confirmed case and 15 deaths, as at 29th March 2016, when the public heath emergency of international concern related to the disease in West Africa was lifted [57].

Continuing public heath awareness campaigns by various governments with support from the World Health Organisation, improvement in the public heath facilities and services and improvement in the living standard have led to reduction in the spread of some of these diseases from areas where they were once known to be endemic, especially in the developing countries where these efforts are needed to counter the high level of superstition owing to the low level of literacy and high level of poverty, which are contributory factors for endemicity of diseases [3].

However, the situation in the African continent is an example of the above scenarios. For example tuberclosis, cholera, smallpox, and other parasitic infections like malaria, schistosomiasis, filariasis, hookworm and trachoma are still endemic in some of these areas of the world [3]. In some of these areas, people are subjected to multiple infections owing to endemicity of two or more infections [3]. This could be in tens of millions, such as leprosy or onchocerciasis, making their total eradication unrealisable at present.

From this discussion, we see that some of these diseases still have high prevalence rate in some areas of the world and so pose formidable challenges to public health authorities. Improving our understanding of their transmission patterns in order to design appropriate intervention therapies that can lead to reduction in their rate of spread in communities where they are known to be endemic is necessary.

1.3 Introduction to the thesis.

Chapter 1 contains an overview and so summarises the work done, history of infectious diseases, their spread and impacts on people living in different parts of the world. We relate our discussions from the past to the present on epidemiology of infectious diseases with focus on the SIR household epidemic model and also arguing the need for carrying out this work. The chapter also contains literature on branching processes, final size probabilities, misclassification of household epidemic data and inference on parameters.

In chapter 2, we examined the SIR household epidemic model, its household structure, branching process for the SIR household epidemic, the community based SIR household epidemic with temporary and permanent immunity of [1], the threshold parameter and its properties in the face of varying local and global infection rates using simulation studies. We also discussed the mean final size of household epidemic, global epidemic and maximum likelihood estimation of the two dimensional model.

In chapter 3, we studied the properties of some functions of the stochastic SIR household epidemic in the face of increasing and decreasing local infection rate, for example the mean final size of household epidemic, threshold parameter and proportion of the initial susceptibles infected.

In chapter 4, we discussed the procedures of fitting the SIR household epidemic model to two dimensional household epidemic data (data from two dimensional model), using the assumption of independence of epidemic in the households and maximum likelihood algorithm in [1]. We compared published results in [1, 9] with those from our program and confirmed them to be the same. Matlab programs to implement the procedures of estimation of the model parameters are discussed with examples using simulation studies, for some choices of theoretical parameters, [1] household structure and minimum epidemic threshold. We also examined the influence of inappropriate choices of the minimum epidemic threshold on the number infected in the households and the precision of the estimates of the model.

In chapter 5, we developed the theoretical foundation leading to four dimensional final size epidemic data and discussed the method of estimation of its parameters. Using simulation

studies, we explored the estimates of the parameters of the three models along the vertical and horizontal axes of the misclassification probabilities region $[0, 0.5)$ and also along the diagonals (slicing through the diagonals of the misclassification probabilities region). We then computed and plotted the root mean square error of the estimates for the three models and presented tables showing the performance of each model and regions of precision of their estimates for given misclassification probabilities.

We also discussed the three dimensional model, which is a particular case of the four dimensional model. Using simulation studies, we compared their estimates with those of the two and four dimensional models and explored their root mean square error for a range of $\varepsilon \in [0, 0.1)$. Table of precision of the estimates of the three models is presented.

In chapter 6, we discussed the procedures of fitting the models to the final size epidemic data using the Pearson Chi-square goodness of fit test and the Kolmogorov-Smirnov goodness of fit test. Using simulation studies and estimation procedures in chapters 5, we fitted the three models to the final size epidemic data, presented their density histograms, plotted the empirical cumulative distribution functions and the cumulative of the hypothesized distribution functions. We also presented tables of mean and variance of the Pearson chi-square test.

We further explored the estimates of three models and their Pearson Chi-square goodness of fit statistics of the model on the four dimensional final size epidemic data, along the diagonals, vertical and horizontal axes of the misclassification probabilities region, $[0, 0.5)$ and plotted the mean and variance of the Pearson Chi-square goodness of fit statistics and the proportion of the simulations rejected from the Pearson Chi-square goodness of fit test at the upper 5% point.

In chapter 7, we employed the Chi-square difference and Kolmogorov goodness of fit tests to examine the model that best fits the final size epidemic data, using simulation studies. We achieved this by computing the Pearson Chi-square goodness of fit statistics and those of their difference statistics (Chi-square difference statistic) for the three models. We then plotted the density histograms of the Chi-square difference statistic with those of their theoretical distributions and the empirical cumulative distributions function with the corresponding

cumulative of the hypothesized distribution.

Also using chi-square difference test, we analysed [1] influenza data those of [28] and identified which model fits significantly better to the epidemic data.

In chapter 8, using simulation studies with Gamma$(k, 4.1/k)$, $k = 1, 2, 5$ infectious period distribution, we examined the effects of misspecification on the estimate of the parameters in the presence of both no misclassification and misclassification of the final size epidemic data. We estimate the epidemic data with a different infectious period distribution from that used in their simualtions. The estimates of the parameters are plotted and tables of mean standard deviation and root mean square error are presented.

In chapter 9, we summarised and discussed our results and their limitations and also provided suggestions for possible extension.

1.4 Background of study.

1.4.1 Empirical Approach to the study of Infectious Disease.

The study of human diseases can be traced to the ancient Greeks e.g. the Epidemics of Hippocrates between $459-377$ B.C [17]. John Graunt $1620-1674$ and William Petty $1623-1687$ made useful contributions through their weekly publication of the London Bills of mortality [17,30]. These were weekly records of London Parishes listing causes and number of death from infectious diseases in Parishes [30], without using any mathematical sophistication or hypothesis on the spread of infections.

Their works set the pace for the development of medical statistics [30]. Fracastorius in 1546 postulated a living principle of contagion, on how disease spreads from person to person, while Daniel Bernoulli in 1760 published his work on variolation of smallpox, in which he showed that inoculation with live virus from patients with a mild case of smallpox confers immunity against the disease [17, 30].

It was not until the work of Pasteur $1822-1895$ which established the link between germs and diseases, in which it was found that boiling liquid destroy germs and the work of Koch between $1843-1910$ who discovered how each type of germ causes a specific disease [2] that

substantial progress began to be recorded in bacteriology science [17]. These results led to corresponding progress in mathematical theories of infectious diseases against earlier empirical descriptions [17].

Definition 1. *The infectious Period of an infected individual is the period during which an infected individual can transmit the disease to susceptible individuals through contacts. It is denoted by* T_I *([9]).*

1.4.2 Work on stochastic epidemic modelling.

The first pioneering work on Mathematical epidemic modelling was proposed by [44] as a continuous time infection model describing spread of an SIR infectious disease in a population of homogeneous mixing individuals. The chance of new infection in short interval of time is assumed to be proportional to both the number of susceptibles and infectives and the length of the interval [2]. An individual is infectious from the moment he receives infectious particles until the moment he dies, recovers or is isolated [17]. Further mathematical theory of epidemics was developed by [38]. However, [34] proposed a probabilistic model of transmission of infectious diseases along the lines of [38, 44] models, in which they assumed that infective and incubation periods are constants. In this model, starting with a single infective in a closed group, new cases will occur in a series of generations. The cases occurring have a binomial distribution, depending on the number of susceptibles and infectives present in the previous generation.

This leads to a chain of binomial distributions [34] in which the distribution of the total number of infectives per household is calculated. In 1928 Lowell J. Reed and Wade Hampton Frost were already discussing and teaching this idea at the Johns Hopkins University in the United States [17, 30].

Also [20] proposed the stochastic version of [38] deterministic epidemic model, referred to as the general stochastic epidemic, which he solved by constructing partial differential equations for the probability generating function of the number of susceptible and infective individuals at any instant [32].

However meaningful explicit solutions could not found because of the non-linear nature of

the transition probabilities, [56] showed that for general case of [18,20], the probability distribution of the ultimate number of infected individuals in [18] may be obtained from solution of set of singly recurrent relations and for large population size, an expression equivalent to [38] threshold theorem is derived. While [33], proposed a Laplace inversion approach which only gives solution for small population sizes (eg. $N = 1, 2, 3$) and the method becomes cumbersome to handle when the population size becomes large, [17,54] proposed various methods to address the non-linear partial differential equations [26].

In 1968 Becker considered some departure from homogeneous mixing assumptions [17], while [21] provided classical results and other features of the deterministic and stochastic models for recurrent epidemics, like the extinction phenomenon which is only peculiar to the stochastic model. Less recursive solution compared to those of [33,54] for the general stochastic epidemic was proposed by [26]. Also [5], provided results for convergence of the general stochastic epidemic to the birth and death process, by constructing a sequence of general stochastic epidemics indexed by the initial number of susceptibles from a time homogeneous birth and death process.

Using a two-type version of a model by [21], while [41] studied the effects of type heterogeneities on the long time behaviour of the models for endemic diseases.

A unified approach to the distribution of the total size and total area under the trajectory of infection (total person time units of infection during the course of the epidemic) was proposed by [6], in which the author showed that if the two assumptions of the general stochastic epidemic which are, (i) Infectious period is exponentially distributed (ii) Population mixes homogeneously, are relaxed then the spread of the epidemic might not follow the SIR epidemic and therefore presented a unified approach to overcome these problems. Results on the convergence of the general stochastic epidemic by [31] were obtained by [7], using coupling arguments and with generalisation to multipopulation epidemics.

Many results on general stochastic epidemic models have been presented by different research workers but the one mostly related to our work is those of [23,55]. They considered a discrete time epidemic among a population partitioned into households [9,11]. The spread of infection within each household is independent and follows a specified distribution [11], such

that infected individuals within the households infect new susceptible households, creating branching process phenomena [9].

An extension of the general stochastic epidemic is given by [9] by assuming a closed and finite population structured into households, each made of susceptibles, infectives and removed individuals, with homogeneous mixing between susceptibles and infectives, independently and at random at two levels, (locally and globally) within the households and individuals from different households, at the points of a homogeneous Poisson processes having rates, λ_L and $\frac{\lambda_G}{N}$ respectively [9], where N is the total population size, λ_G is the total rate that a given infective makes global contacts [6, 9].

In this model any susceptible contacted will immediately become infectious (since we assume that there is no latency for the disease) for period T_I, referred to as the infectious period after which it is removed (died or isolated or immune) at the end of the infectious period. The infectious period of each infective is assumed to be independent and identically distributed according to the random variable T_I which is arbitrary but must be specified [6, 9]. The Poisson processes describing contacts and the infectious period are assumed to be mutually independent.

1.5 Misclassification of household epidemic data.

Measurement error occurs when in an analysis the real variable is unavailable and replaced by its surrogate. Such analyses are often referred to as naive [35]. For example in a regression analysis with explanatory variable X and response Y either of the variables can be subject to mismeasurement. Suppose Y is subject to mismeasurement so that Y^* is observed in place of Y, where Y^* is obtained by adding noise to Y, independent of the true explanatory variable X, where the noise is assumed to be normally distributed with mean 0 and variance 1. Then [35] showed that adding noise to the response variable Y does not shift the estimated regression slope of the line, but rather increases the uncertainty (standard error) about the relationship between the variables.

On the other hand if it is the explanatory variable X that is subject to mismeasurement,

then adding noise to the explanatory variable imparts a bias.

Mismeasurement of the explanatory variable is often associated with a flattening or attenuation in the strength of the association between the explanatory and response variables. This also carries over to categorical variables as in [35].

On categorical data, mismeasurement occurs when the actual and recorded categories for subjects differ. In that case, the surrogate variable cannot be expressed as sum of the true variable plus a noise variable, rather they are expressed in terms of classification probabilities often referred to as misclassification probabilities.

For example, let X be a random variable representing individual health status and $X = 1$ is the event that individual is observed correctly as having a particular health status with a corresponding probability $P(X = 1) = 1 - \varepsilon$.

Suppose individuals are not observed correctly, instead a surrogate X^* is observed in place of the true value X. Then the misclassification error model of X^* given the true value of X is written as, $P_{x^*\,|\,x} = P(X^* = x^*\,|\,X = x)$. Here, $P_{1\,|\,1} = P(X^* = 1\,|\,X = 1)$ and $P_{0\,|\,0} = P(X^* = 0\,|\,X = 0)$ are the sensitivity and specificity.

The sensitivity and specificity parameters are used to measure the magnitude of the misclassification [35] and defined respectively as the probability of correctly classifying an infective as one with the disease while specificity is the probability of correctly classifying a susceptible as one without the disease.

The relation between the surrogates and the exact variables can be seen in terms of the misclassification probabilities, $P_{1\,|\,0} = (1 - P_{0\,|\,0})$ and $P_{0\,|\,1} = (1 - P_{1\,|\,1})$. Here $P_{1\,|\,0}$ is the false positive misclassification probability while $P_{0\,|\,1}$ is the false negative misclassification probabilities.

Misclassification of binary variable is represented by a 2×2 matrix which is a function of the misclassification probabilities. We denote it by $P(\varepsilon_{FN}, \varepsilon_{FP})$ and has the form,

$$P(\varepsilon_{FN}, \varepsilon_{FP}) = \begin{pmatrix} P_{0\,|\,0} & P_{0\,|\,1} \\ P_{1\,|\,0} & P_{1\,|\,1} \end{pmatrix}$$

where $P_{0|1} = \varepsilon_{FN}$ and $P_{1|0} = \varepsilon_{FP}$. Hence $P_{0|0} = 1 - \varepsilon_{FN}$ and $P_{1|1} = 1 - \varepsilon_{FP}$ are the true

negative and positive misclassification probabilities. If the misclassification probabilities are the same, then the epidemic model is three dimensional with the corresponding misclassification probabilities,

$$P(\varepsilon) = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$$

If the probability of classifying an infective as a susceptive is different from the probability of classifying a susceptible as an infective then the model will be referred to as four dimensional having matrix of misclassification probabilities,

$$P(\varepsilon) = \begin{pmatrix} 1-\varepsilon_{FN} & \varepsilon_{FN} \\ \varepsilon_{FP} & 1-\varepsilon_{FP} \end{pmatrix}$$

Here four refers to number of parameters to be estimated which are the two different misclassification probabilities, the local infection rate and the probability of avoiding infection from the population. These concepts are further discussed in section 1.10

1.5.1 Epidemic modelling in the presence of misclassification.

Parameter estimation in the presence of noise is often a problem in stochastic model fitting to epidemic data. Not much work has been done in this area and so limited literature is currently available. However, ignoring it will lead to biased estimates of the parameters and hence model that does not fit.

Therefore such models are unreliable for use in projections. Hence there is the need to take cognisance of these errors in our inferences by using appropriate models that adjust for the errors in the data and still give precise estimates. This is what this work intends to achieve.

Errors in the data may be due to number of factors, which could include use of defective measurement devices or inaccurate diagnostic procedures (for disease status), leading to wrong classification of individuals as positive when not or wrongly classified as negative when positive.

1.5.2 Literature on modelling misclassified finite count data.

Many methods have been proposed in the biomedical literature by research workers, especially on regression modelling, when both the exposure and response variables [35] are categorical with the classification error expressed in terms of classification probabilities.

Notable among them is the work of [40] who proposed correction methods for misclassified finite count data in their dental caries research studies. In their work in [40], they assumed Y to be the true finite count with range $\{0, 1, \ldots, K\}$ and binary scores $\{Z_1, Z_2, \ldots, Z_K\}$ which make up the count, i.e. $Y = \sum_{k=1}^{K} Z_k$. They assumed the possible observed corrupted counts to be Y^* and Z^* so that $Y^* = \sum_{k=1}^{K} Z_k^*$, also for $(r,\, s = 0, 1, \ldots, K)$, let $\pi_{r\,s} = P(Y^* = r \,|\, Y = s)$ such that $\sum_{r=0}^{K} \pi_{r\,s} = 1$, be the misclassification probabilities represented by the vector $\pi_s = (\pi_{0\,s}, \pi_{1\,s}, \ldots, \pi_{K\,s})^T$.

Assuming that $n_{r\,s}$ is the number of individuals with $Y = s$ and $Y^* = r$, and the misclassification process is non-differential [40], i.e. that the misclassification probabilities are constant over individuals then we get a $(K+1) \times (K+1)$ classification table. Also assuming independence of the subjects, the sth column, $\mathbf{n}_s$ of the misclassification table with $n_{r\,s}$ follows a multinomial distribution, $\mathbf{n}_s \sim \text{Multinomial}(n_s,\, \pi_s)$ with the Multinomial estimates of $\hat{\pi}_{r\,s} = {}^{\pi_{r\,s}}/_{\sum_{r=0}^{K} n_{r\,s}}$ and variance var $= {}^{\pi_{r\,s}(1-\pi_{r\,s})}/_{\sum_{r=0}^{K} n_{r\,s}}$ [35].

The process is such that in order to obtain the count Y, it is required that we score the binary indicators, $Z_k, k = 1, \ldots, K$ which are the gold standard (true values), and Z_k^* (examiner) are available in a validation study. The number of individuals with r examiners and s gold standards, $Z_k^* = r$ and $Z_k = s$, were defined as $n_{k,r,s}$. The sensitivity $\alpha_k = P(Z_k^* = 1 \,|\, Z_k = 1)$ and specificity of the binary indicators $\beta_k = P(Z_K^* = 0 \,|\, Z_k = 0)$ were then obtained from the corresponding 2×2 classification table with entries $n_{k\,r\,s}$ for $k = 1, 2, \ldots, K$ as,

$$\alpha_k = \frac{n_{k,1,1}}{n_{k,0,1} + n_{k,1,1}},$$
$$\beta_k = \frac{n_{k,0,0}}{n_{k,0,0} + n_{k,1,0}}.$$

By assuming independence for the binary indicators with scoring behaviour independent of k such that $\alpha_k = \alpha_Z$ and $\beta_k = \beta_Z$ for all $\{k = 1, 2, \ldots, K\}$ and the subject (non-differential

misclassification), [40] proposed a double binomial approach (DB) which expresses the distribution of (examiner) r given that the true is s (gold standard), as the sum of two independent binomial distributions $\text{Bin}(s, \alpha_Z)$ and $\text{Bin}(K-s, 1-\beta_Z)$, where the maximum likelihood estimates of α_Z and β_Z are obtained from the corresponding 2×2 classification tables given as,

$$\hat{\alpha}_Z = \frac{\sum_{k=1}^{K} n_{k,1,1}}{\sum_{k=1}^{K} \{n_{k,0,1} + n_{k,1,1}\}},$$
$$\hat{\beta}_Z = \frac{\sum_{k=1}^{K} n_{k,0,0}}{\sum_{k=1}^{K} \{n_{k,0,0} + n_{k,1,0}\}}.$$

The probabilities for the misclassification table for Y are derived from the misclassification table for Z_k, $k = 1, 2 \ldots, K$ as,

$$\pi_{r\,s} = \sum_{m=M_0}^{M_1} \binom{s}{m} \alpha_Z^m (1-\alpha_Z)^{(s-m)} \binom{K-s}{r-m} (1-\beta_Z)^{(r-m)} \beta_Z^{(K-s-r+m)}. \tag{1.5.1}$$

Where $M_0 = \max(r-(K-s), 0)$, $M_1 = \min(r, s)$ [40].

The first binomial distribution expresses the probability that the examiner scores m teeth in the caries research as decayed from s teeth that the gold standard was scored decayed [40], while the second binomial distribution expresses the probability that the examiner scores $(r-m)$ teeth as decayed from the $(K-s)$ teeth that the gold standard scored not decayed.

As an alternative to $\hat{\alpha}_Z$ and $\hat{\beta}_Z$, [35] observed that one can estimate the sensitivity and specificity directly from the Multinomial model, where $\pi_{r\,s}$ is given by equation (1.5.1). They observed that such estimates are often close to the Multinomial estimates especially if the size of the validation studies is large [35].

Since some of the assumptions of the multinomial model might not hold in practice, all types of extension of the binomial models could be employed. Hence [35] proposed DB model

extension,

$$P(Z_1^*, Z_2^*, \ldots, Z_K^* \mid Z_1, Z_2, \ldots, Z_k) = \prod_{k=1}^{K} P(Z_k^* \mid Z_k),$$

where $\alpha_Z = \alpha_Z(f(Z_1 \ldots, Z_K))$ and $\beta_Z = \beta_Z(f(Z_1, \ldots, Z_K))$ and that in dental research when $f(Z_1, \ldots, Z_K) = \sum_{k=1}^{K} Z_k$, sensitivity and specificity depends on the number of caries in the mouth [35].

A method of evaluating the effect of misclassification on the estimation of a disease relative risk from retrospective studies was proposed by [19]. They assumed a population classified according to the presence or absence of each of two traits [19] and constructed a 2×2 matrix of their joint probabilities and that of their conditional probabilities of misclassification [19]. Then they found that relative risk can be calculated in terms of the entries in the matrix of the misclassification probabilities, if estimates of the false positive and negative rates for the method are available.

Using maximum likelihood estimation method, [43] examined the degree of estimation error of household and community transmission parameters from influenza infection data due to misclassification of infectives and susceptibles in a stochastic simulation model [43]. The expected numbers of detected infectives at different levels of sensitivity and specificity were simulated for the serological tests used. It was found that the maximum likelihood estimator for the household transmission parameter is precise.

1.6 The stochastic SIR epidemic model.

In this section, we discussed the stochastic version of [38, 44] deterministic SIR epidemic model. It is a continuous-time stochastic process defined on a closed and finite population in which the population is partitioned into susceptibles, infectives and removed individuals denoted by $\{S(t), I(t), R(t), \}$, for $t \geq 0$. A susceptible individual is one that can be infected with the disease under discussion, an infective is one that has the disease and can transfer it

to the susceptibles through contacts, while a removed individual is one that has recovered and is immune or has died from the disease. Such a person makes no contribution to the disease transmission process.

The model assumes random mixing (homogeneous mixing) between individuals in the population which occurs independently and at random at the points of a Poisson process having rate λ/n [2], where n is the initial number of susceptibles.

The infectious periods of different infectives are independently and identically distributed according to the random variable T_I having arbitrary but specified distribution [2].

All the Poisson processes are assumed to be independent of each other and of the infectious period of the disease. If a susceptible individual is contacted by an infective during the infectious period then it will immediately become infected and infectious. The newly infected individual will also continue the transmission process to other susceptibles in the population. The epidemic ceases as soon as there are no more infectious individuals in the population. The case with exponential distributed infectious period, which is referred to as the general stochastic epidemic was proposed by [20].

However, the SIR epidemic can be made to have multiple types of individuals [12], $\mathbb{K} = 1, 2, \ldots, k$ with population of susceptible, infective and removed individuals of types k as $\{S_k(t), I_k(t), R_k(t)\}$ at time $t \geq 0$. An individual of type k will have infectious period distributed according to $T_{I,k}$ and also makes contact with an r susceptible at the points of Poisson process having rate $\alpha_{r,k}$. The contact rates can then be stored in a $\mathbb{K} \times \mathbb{K}$ matrix A. If $\mathbb{K} = 1$, the population is made of one type of individual, then we recover the SIR single type population epidemic discussed earlier.

1.7 Branching process.

Branching processes are stochastic processes used in analysing changes in population over time, e.g. in approximating the size of epidemics in the early stages. Branching processes were first proposed by Francis Galton and Reverend H. W. Watson in 1874 in their study of extinction of family names [30].

It is based on the assumption that each individual is associated with life-length, often referred to as generation time or individual life-span [2], and at the end of the generation time produces a random number of offspring independent of the rest of the population. Several of these processes have been developed and used in approximating epidemics [22] . Among them are the one-type and multi-type Galton-Watson branching processes which assume a fixed length for the generation time and that at the end of the generation time produce random numbers of offspring in line with the above definition [27]. The Bellman-Harris branching process assumes that individuals have independent and arbitrary distributed generation time and produce random numbers of offspring independently only at the end of the generation time.

The Bienayme-Galton-Watson processes simply called BGW-processes, only takes account of successive generations of offspring in a discrete formulation [47]. Crump-Mode-Jagers Processes, simply referred to as CMJ-processes, are generalised age-dependent extensions of the BGW-process. They were independently proposed by Crump, Mode and Jagers to accommodate cases of individuals producing offspring at random points throughout their lifetimes [47].

The CMJ processes assumes that individuals have independent and arbitrarily distributed generation time. Each individual produces offspring according to a counting process throughout their generation time. Different individuals follow the same counting process. The generation time and the counting process are independent.

A Crump-Mode-Jagers process is sometimes used to approximate epidemics in their early stages because of the similarity of its assumptions with the stochastic SIR epidemic process. For example the generation time in the CMJ process corresponds to the infectious period in the epidemic process, the assumption of arbitrary distributed generation time also agrees with that of the infectious period in an epidemic process. The way in which individuals produce new offspring at random at points of counting process in the branching process corresponds to homogeneous contacts at points of Poisson process in the epidemic process.

Finally, the assumption of independence of contact processes and the distribution of the generation time in the branching process and infectious period in epidemic process agree.

1.8 Convergence of the general stochastic epidemic.

Let $\{Y_r(t),\, t \geq 0\}, r = 1, 2 \ldots,$ be the number of infectives in the rth epidemic, and the sequence $\{Y(t),\, t \geq 0\}$ be the number of individuals alive in the continuous time branching process denoted $E_a(\lambda, I)$. Here λ is the birth rate and I, individuals life span in the branching process assumed independent but identically distributed and a is the number of initial ancestors. Then [2] in line with [5] showed that the sequence of the epidemic processes converges to the associated branching process using the following definitions.

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be the probability space with individual life histories $\mathfrak{H}_{(\mathfrak{a}-1)}, \mathfrak{H}_{(\mathfrak{a}-2)}, \ldots,$ where $\mathfrak{H}_{\mathfrak{i}}$ is list containing the life span of the ith individuals together with the time points at which this individuals gives birth [2], a is the number of initial infectives.

Also let $\{U_i,\, i \geq 0\}$ be a sequence of independent and identically distributed random variables defined on the above probability space, each uniformly distributed on $(0, 1)$ and $E_{n,a}(\lambda, T_I)$, $n \geq 1$, be a sequence of epidemic processes with a initial infectives, infection rate of λ and infective period, T_I. Now fix the number of susceptibles and label them as $1, 2, \ldots, n$. We see that the initial ancestors in the branching process corresponds to the initial infectives in the epidemic process.

Contact occurs in the epidemic process whenever a birth occurs [2] in the branching process and the individual who is contacted at the i^{th} contact has label, $d_i = [nU_i] + 1$ [2]; where $[x]$ is the largest integer $\leq x$. If the contacted individual is still susceptible then she will become infected in the epidemic process, otherwise she and all of her descendants in the branching process (often referred to as ghosts) are ignored in the epidemic process [2].

The death of non-ghost individuals in the branching process agrees with removal in the epidemic process. Thus, the processes Y_n and Y agree until time T_n of the first ghost.

The number of births in the branching process during a fixed time interval $[0, t_0]$ is finite almost surely [10]. It is observed by [2] that any finite number of labels d_i will be distinct with a probability tending to 1 as $n \to \infty$, that

$$P(T_n > t_0) = 1, \text{ as } n \to \infty.$$

An important threshold theorem which determines the nature of an outbreak of the stochastic SIR epidemic in large population is provided by [10] and also reported in [2]. We will state the theorem without proof as it is in [2]. The proof can be found in [10].

Theorem 1. *Consider a sequence of epidemic processes $E_{n,a}(\lambda, T_I)$, $n \geq 1$. Also denote by $Y_n(t)$ the number of infectives at time t, $t \geq 0$. Then for each fixed t_0 $Y_n(t_0) \to Y(t_0)$ almost surely, where $\{Y(t);\, t \geq 0\}$ is the process describing the number of individuals alive in the branching process $E_a(\lambda, I)$.*

If $\lambda\ell \leq 1$ then Y becomes extinct with probability 1. On the other hand , if $\lambda\ell > 1$ then Y becomes extinct with probability q^a, where q is the smallest root of the equation $\phi(\lambda(1-\theta)) = \theta$, or explodes with probability $1-q^a$, $\ell = E(T_I)$ is the mean infectious period, while $\phi(\lambda(1-\theta))$ is the moment generating function of the number of individuals infected whose smallest solution is q.

In this theorem, $\lambda\ell$ is the mean number of individuals infected.

This theorem shows that the threshold parameter $R_0 = \lambda\ell$ determines the probability of an epidemic being a minor or a major epidemic.

1.9 Final size household epidemic data.

Final size data are observational data on the size of outbreaks of epidemic in households. It is a collection of epidemic data classified by size of households and corresponding number infected. It is often represented in matrix form with rows representing household types and column entries the infection sizes for the corresponding household type as in tables 1.1 and 1.2.

The entries in table 1.1 are described as the number of households of the given type with the given infection size. For example, all the column entries corresponding to households of sizes one can be read as, number of households of size one with zero infectives, and number of households of size one with one infection respectively.

Also the entries corresponding to household of size two in the second row of the matrix are read as, number of households with zero infectives, followed by number of households

Household Size	Number Infected in Household					
	0	1	2	3	4	5
1	$n_{1,0}$	$n_{1,1}$	-	-	-	-
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$	-	-	-
3	$n_{3,0}$	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	-	-
4	$n_{4,0}$	$n_{4,1}$	$n_{4,2}$	$n_{4,3}$	$n_{4,4}$	-
5	$n_{5,0}$	$n_{5,1}$	$n_{5,2}$	$n_{3,3}$	$n_{5,4}$	$n_{5,5}$

Table 1.1: Final size household epidemic data.

with one infective and finally number of households with two infectives, up to the maximum household size.

An example of final size household epidemic data is that of [1] which was obtained in their study of transmission of influenza A(H3N2) in Tecumseh, Michigan USA. It is presented in table 1.2, with entries having the same meaning as that of table 1.1. For example $n_{1,0} = 110$, is the households of size one with zero infectives and $n_{1,1} = 23$ is the households of size one with one infective etc. We will analyse this dataset in section 4.4 of chapters 4, section 6.14 of 6 and section 7.13 of chapter 7 respectively.

Household size.	Num. Inf. in Houshold.					
	0	1	2	3	4	5
1	110	23	-	-	-	-
2	149	27	13	-	-	-
3	72	23	6	7	-	-
4	60	20	16	8	2	-
5	13	9	5	2	1	1

Table 1.2: Each coefficient represents number of households of a particular size with number of infectives by the end of the epidemic.

1.10 Dimensionality of household epidemic data.

In this thesis, it is assumed that when individuals are correctly classified; then data from such classification will be referred to as two dimensional final size epidemic data, the corresponding dimensional model as two dimensional model while the associated numerical optimisation for

the estimation of its parameters as two dimensional numerical optimisation. Here, two refers to the number of parameters to be estimated from the model. Similarly the three dimensional household epidemic data means three parameters are to be estimated from the model, where the parameters to be estimated are the misclassification probability (where the false negative misclassification probability and the false positive misclassification probability are assumed to be the same), the local infection rate and the probability of avoiding infection from the households.

Sometimes these misclassification probabilities may be different from each other. If that occurs, the final size data is referred to as four dimensional final size epidemic data and the associated numerical optimisation used in the estimation as the four dimensional numerical optimisation. This means that four parameters are to be estimated from the final size data and the associated model will be referred to as four dimensional model.

There may be need to fit the three models to the same final size data in order to compare which of them is significantly better on the final size data, especially when misclassification error in the final size epidemic is known to have occured. We accomplished this using simulation studies in chapters 5, 6 and 7 respectively.

1.11 The Gontcharoff polynomial.

The Gontcharoff polynomial was first proposed by Gontcharoff in 1937 but were not extensively utilised by research workers until developments in stochastic modelling [9, 16, 39]. As it becomes necessary to explore simpler mathematical methods for solutions to some epidemiological problems, like finding expressions for final size distribution and the total size infection as demonstrated by [9, 39].

The Gontcharoff polynomial attached to the sequence of real numbers, $U = u_0, u_1, \ldots,$ is defined in [9, 39] as $G_0(x \mid U), G_1(x \mid U), \ldots,$ and obtained recursively as,

$$\sum_{j=0}^{i} \frac{u_j^{i-j}}{(i-j)!} G_j(x \mid U) = \frac{x^i}{i!},$$

where for $i = 1, 2, \ldots$ the polynomial $G_i(x \,|\, U)$ satisfies the integral representation.

$$G_i(x \,|\, U) = \int_{u_0}^{x} \int_{u_1}^{\xi_0} \cdots \int_{u_{i-1}}^{\xi_{i-2}} \mathrm{d}\xi_0 \mathrm{d}\xi_1 \ldots \mathrm{d}\xi_{\mathrm{i}-1}.$$

Also for $0 \leq k \leq j$, the kth derivative of the Gontcharoff polynomial, $G_i^k(x \,|\, U)$ is defined by,

$$G_i^{(k)}(x \,|\, U) = G_{i-k}(x \,|\, E^k U),$$

where the operator $E^k U$ generates the sequence, $u_k, u_{k+1}, u_{k+2}, \ldots$ and $G_i^k(x \,|\, U) = 0$ if $k > i$ [9, 16].

This approach was employed by [39] to derive equations for the final size distribution and the total size of infection. Further discussion can be found in section 2.8.

Chapter 2

The stochastic SIR household epidemic model.

In this chapter, we examined the properties of the SIR household epidemic model. These include its household structure, contact processes, branching process approximation of the epidemic, community based version of the SIR household epidemic model, the threshold parameter, the mean final size, global epidemic and maximum likelihood estimation of the parameters of the model.

2.1 Introduction.

Early pioneering work on modelling of infectious diseases in populations structured into households, can be traced to the work of [52], which considered a continuous time deterministic simple epidemic, without removal of infectives, the so called SI epidemic model, in a large population. The work of [23, 55], which are relevant to our work, considered a discrete time epidemic among a population of households in which the spread within households followed independent and random but specified processes and at each time point infected individuals independently and at random infect a number of new susceptible households, creating a branching process scenario of the epidemic process.

2.2 Household structure.

The stochastic SIR household epidemic model of [6,9] is based on population of size $N \in \mathbb{Z}_+$, partitioned into households sizes $n \in \mathbb{Z}_+$ with the proportion of households of size n denoted by [9] as

$$\alpha_n = \frac{M_n}{M}, \tag{2.2.1}$$

where M_n is the number of households of size n and M is the total number of households. Here, $N = \sum_{n=1}^{\infty} nM_n$.

Definition 2. *Contacts between susceptibles in the households and infectives from other households are referred to as global contacts, while contacts between susceptibles and infectives within the households are called local contacts.*

Let $\tilde{\alpha}_n$ be the probability that a global contact is with an individual residing in a household of size n [9, 11], then

$$\tilde{\alpha}_n = nM_n/N.$$

2.3 Two level mixing epidemic model.

The stochastic SIR household epidemic model of [9] sometimes referred to as the two level mixing model is well discussed in [9,11]. It is a generalisation of the stochastic SIR epidemic, designed to study disease outbreaks in a population divided into households, identify number of individuals infected, their distributions and also identify possible vaccination strategies for their control.

The population is assumed to be closed and finite (without birth, or death), structured into small groups or households. Each household is made of susceptibles, infectives and removed individuals, with contacts between susceptible and infective individuals occurring at two levels, within and between the households (locally and globally) independently and at random, at points of homogeneous Poisson processes having rates, λ_L and $\frac{\lambda_G}{N}$ respectively as discussed in [9], where N is the total population of individuals in the households, λ_G

is the total rate that a given infective makes global contacts and λ_L is the local contact rate (contacts between individuals in the households) as in [9]. Any individual contacted if susceptible will immediately become infectious, for period (as there is no latency for the disease) T_I, referred to as the infectious period after which the individual is removed (died or quarantined or immune) at the end of the infectious period, as it no longer plays any part in the epidemic. We assumed no disease latency, as the distribution of the final size of the epidemic is invariant to general assumptions concerning the latency period [9]. The infectious period of each infective is assumed to be independent and identically distributed according to the random variable T_I which is assumed to be arbitrary but must be specified in line with [9]. The Poisson processes describing contacts and the infectious period are assumed to be mutually independent [9, 11].

However, [14] proposed a general stochastic model with two levels mixing with household, overlapping groups and great circle models as special cases, where in the household model, mixing occur within the households and a much smaller lower rate within the population [14].

Here, an individual $i \in N$ (where N is the population size) who is infectious is assumed to make local contacts with an individual $j \in N - \{i\}$ at the points of a homogeneous Poisson process with rate $\lambda^L_{i,j}$. Where $\lambda^L_{i,j} = \lambda_L$, if i and j individuals are from the same household, otherwise is 0 [14]. It also makes global contacts with individuals chosen uniformly from the population at the point of a homogeneous Poisson process with rate, λ_G/N, where λ_G is the individual to individual global contact rate [14].

In the overlapping case, the population is partitioned in several ways with uniform mixing between individuals within the partition [14] and also global mixing with the population. While in the case of the great circle model, the population is assumed to be equally spaced around a circle [14] such that local infection occur between the nearest-neighbour.

Similarly, [50] proposed two levels stochastic SIR model with changing group of contacts during the day in contrast to [9, 14], where each day is divided into morning and night with length of the morning period at the start of each day, $0 < \tau \leq 1$, and night period, $1 - \tau$, [50].

Where, contacts made by individuals in the population are assumed to depend on the time of the day, morning and night. That is, during the morning the whole population is

mixing together and then individuals return to their homes (households) at night [50]. Thus, infectious individuals make contact with individuals in the population during the morning and at night can only infect their household members. Under this settings and using branching process approximation for the epidemic, [50] discussed the epidemic at the initial stages and the probability of a major epidemic outbreak.

Using Monte Carlo simulations, [15] studied the initial behaviour and the final outcome of a SIR network model with two level mixing (local and global) under weak constraints on the prescribed degree distribution and showed that the asymptotic results provide a good approximation even for moderately small population size, N [15]. Here, contacts are with individuals in the same network and those in the population [15], where the networks are represented with unidirected random graphs representing the possible individuals he/she is connected to in the network and can therefore infect if he/she is infectious.

2.4 Branching process for two level mixing epidemic model.

If the population is large and the number of initial infectives is small then during the early stages of the epidemic the probability that global contact is with an individual residing in a previously infected household is small [9]. Then [9] showed that the initial stage of the epidemic can be approximated by a branching process in which at time $t = 0$, an initial infective infects susceptible members of its own household and other households.

Those infected form the first generation of infectives. Individuals infected by the first generation of infectives also infect other susceptible members of their households and susceptible individuals in other households. This process of creating new infections locally and globally follows a branching process until the first contact with an infective or removed individual (often referred to as a ghost).

During its infectious period an infective makes global contact with distinct individuals in the households independently and randomly, at the points of a Poisson process having rate λ_G. The total number of global contacts from the household epidemic, R_n, follows a Poisson distribution with random mean, $\lambda_G T_A$, where T_A is the sum of the infectious periods of the

infectives, T_A is also referred to as the severity of the epidemic [9, 11].

Let R be the number of infected households emanating from the household epidemic in line with [11], then R is the offspring random variable for the approximating branching process, in the epidemic process. This is defined as the total number of infected households caused by an infected household throughout the infectious period [11]. Now let $R_* = E(R)$, be the mean number of infected households from an infected household from the household epidemic and $E(\theta^R) = h(\theta)$ be the probability generating function of the offspring variable random R.

Then in line with [9, 47] standard branching process theory, a global epidemic occurs if in the limit, as the number of households, $m \to \infty$, the epidemic infects infinitely many households.

2.5 Community based SIR household epidemic model with temporary immunity.

In [42] the dependency assumption of epidemic assumed in [6, 9, 11] was ignored and instead proposed a community based transmission stochastic SIR transmission model, in which susceptibles in the households are infected from the community and from infectives within their households. They assumed that every susceptible in the household has equal probability of avoiding infection from the community, written as $b_t = 1 - a_t$, where a_t is the probability that a susceptible from a household becomes infected from the community, $t = 0, \ldots, T$ is the time period of infection, and a bounded function, $B = f(b_t)$ defined as the probability that an infective is not infected from the community [42]. A particular case is when $B = b_t$. However, [42] model is limited to cases of infectious diseases which confers temporary immunity.

2.6 Community based SIR household epidemic model with permanent immunity.

Another variant of [42] model is that of [1] which is concerned with spread of infectious diseases that confer permanent immunity. The model allows heterogeneity in contact and

multiple source of infection. The population is stratified according to different group of individuals ($i = 1, \ldots, m$) say, with each individual in exactly one group and susceptible to the infectious disease of interest [1]. They assumed that an epidemic can be started by one or more individuals, a_i, $i = 1, 2, \ldots, m$, becoming infected from a specified source outside the population, similar to the assumption of [6], where a_i, $i = 1, 2, \ldots, m$, are the initial number of infectives in group i. A similar model, with the same assumptions is given by [36] but focuses on design of vaccination studies.

The initial number of susceptible individuals are assumed to be, $\mathbf{N} = (N_1, \ldots, N_m)^{'}$ with the total population size of the susceptibles $N = \sum_{i=1}^{m} N_i$, $i = 1, 2, \ldots, m$. The length of the infectious period of an i infective residing in $k = 1, 2, \ldots, m$ group is assumed to be $T_{i,k}$, with moment generating function, $\phi_i(t) = E(\exp(-tT_{i,k}))$. The progress of the epidemic in each household is independent [1], contrary to [9] which assumed dependency of epidemic between households. Given these assumptions, the epidemic is governed by two parameters, namely extra-population escape probability, defined as the probability that a susceptible of type $i = 1, 2, \ldots m$ escapes infection from outside the population during the course of the epidemic represented by $\mathbf{B} = (\beta_1, \ldots, \beta_m)^{'}$, where each $\beta_i, i = 1, \ldots m$ is the extra-population escape probability for susceptibles of type $i = 1, \ldots, m$. Also the within-population disease transmission, defined as the rate at which a susceptible from group of type i comes in contact with an infective from a group of type k is represented by [1] as $\beta_{i,k}$.

2.6.1 Calculation of the final size probabilities.

Using the assumptions in section 2.6, the triangular equation for the probability of the final size household epidemic assuming the value ω given m groups of different types of individuals is,

$$1 = \sum_{\boldsymbol{\omega}=0}^{\mathbf{j}} \binom{\boldsymbol{j}}{\boldsymbol{\omega}} P^{\mathbf{N}}_{\boldsymbol{\omega}_i \ldots \boldsymbol{\omega}_m} / \boldsymbol{\phi}(\boldsymbol{\beta}^{'}(\mathbf{N} - \mathbf{J}))^{\boldsymbol{\omega}+\boldsymbol{a}} \mathbf{B}^{(\mathbf{N}-\mathbf{J})}, \ \mathbf{j} \geq 0,$$

where $\omega = (\omega_1, \ldots, \omega_m)$, $\boldsymbol{\beta}$ is an $m \times m$ of contact rates, $\mathbf{B}$ is a vector of all the extra-escape population probabilities, while $\mathbf{N}$ is the vector of all the initial susceptibles in the m groups of different types of individuals.

If the number of households approaches infinity in [9], then each given susceptible in each group independently avoids infection from outside the population with the same probability [9]. This is in agreement with the assumption of [1]. Thus under this large population assumption in [9], the ultimate spread of infection within the group has the same distribution as that of the extended model of [1].

Thus, for population with single type of individual, $m = 1$ for which $a = 0$, the final size probabilities satisfy the triangular equation,

$$1 = \sum_{\omega=0}^{j} \binom{j}{\omega} P_{\omega}^{n} / \phi(\lambda_L(n-j))^{\omega} \pi^{n-j}, \;\; j \geq 0, \tag{2.6.1}$$

where λ_L is the local contact rate, π is the probability of avoiding infection from outside the household given in [9] and P_w^n, are the final size probabilities of the epidemic outcomes $w = 0, 1, 2 \ldots, n$ and n is the household size [6, 9, 11].

Rearranging the triangular equation (2.6.1), the final size probability is given by,

$$P_k^n = \left(1 - \sum_{w=0}^{k-1} \frac{\binom{k}{w} P_w^n}{\phi(\lambda_L(n-k))^{w} \pi^{n-k}}\right) \phi(\lambda_L(n-k))^{k} \pi^{n-k}, \;\; k = 0, 1 \ldots n, \tag{2.6.2}$$

where n is the number of the initial susceptibles in the household and $\phi(\lambda_L) = E(\exp(-\lambda_L T_I))$.

Taking into account all the possible ways an individual can become infected, the final size probabilities are given by [6, 9, 11] as,

$$P_{n,i} = \binom{n}{i} P_i^n. \tag{2.6.3}$$

2.7 Threshold parameter.

Using the branching process theory in section 2.4, a global epidemic can occur for the stochastic SIR household epidemic if the threshold parameter, defined as the mean number of infected households generated by a single infected household, $R_* > 1$.

Recall that R_n is the total number of global contacts from the single household epidemic, which is a Poisson distributed random variable, with mean $T_A\lambda_G$. The threshold parameter can be written as,

$$\begin{aligned} R_* = E(R) &= \sum_{n=1}^{\infty} \tilde{\alpha}_n E(R_n), \\ &= \sum_{n=1}^{\infty} \tilde{\alpha}_n \lambda_G E(T_A), \end{aligned}$$

where the distribution of T_A depends on the household size n, $\tilde{\alpha}_n$ is the probability that global contact is with an individual residing in a household of size n [9, 11].

We can write $E(T_A) = E(T_n)E(T_I)$ [11] using the Wald's identity for epidemic [6], where T_n is the number of infected individuals, including the initial infectives, by the single household epidemic, T_I is the infectious period of each infective in the single household epidemic. Since $\mu_n = E(T_n)$, $n = 1, 2 \ldots$, the threshold parameter is simplified in [11] as,

$$R_* = \lambda_G E(T_I) \sum_{n=1}^{\infty} \tilde{\alpha}_n \mu_n. \tag{2.7.1}$$

Since the threshold parameter for single population SIR stochastic model in which the households are all of size one, is $R_0 = \lambda_G T_I$, where $E(T_I)$ is the mean infectious period of the household epidemic, we can express the threshold parameter for the household epidemic as, $R_* = R_0\mu$ [9, 11] where $\mu = \sum_{n=1}^{n} \tilde{\alpha}_n \mu_n$ is a mean amplification owing to internal spread within household. The parameter $\mu_n = \mu_{n-1,1}$ is the mean final size of the household epidemic with $n-1$ initial susceptibles and 1 infective.

2.8 Mean final size of single household epidemic.

The mean final size for an epidemic in a single household with single initial infection can be generalised to the case with a initial infectives and n susceptibles in the household written as $\mu_{n,a}$. One method of computing this function which uses the non-standard family of polynomials introduced in section 1.11 is provided by [9, 39]. However, [11] provided an

alternative method without employing the Gontcharoff polynomial, while [9] also obtained the same result using the joint generating function of the final size and severity of the epidemic. Where the moment generating function of the infectious period, T_I is given by [9, 11] as,

$$\phi(\theta) = E(\exp(-\theta T_I)),\ \theta \geq 0,$$

so that the joint distribution of the final size, T and severity of the epidemic, T_A is the written by [9] as,

$$\phi_{n,a}(s,\theta) = E(s^{n-T}\exp(-\theta T_A)),\ \theta \geq 0,$$

where n is the initial number of susceptibles and a that of the infectives.

The joint moment generating function can be written as,

$$\phi_{n,a}(s,\theta) = \sum_{i=0}^{n} \frac{n!}{(n-i)!}\phi(\theta + \lambda_L i)^{n+a-i} G_i(s \mid U), \tag{2.8.1}$$

where the sequence $U = \phi(\theta + \lambda_L i),\ i = 0, 1, \ldots,$

From the definition of T, let $\mu_{n,a} = E(T)$ be the mean final size of a household epidemic with n susceptibles and a initial infectives. Then differentiating $\phi_{n,a}(s,\theta)$ with respect to s, and setting $s = 1$, $\theta = 0$, [9] shows that the mean final size of the household epidemic can be written as,

$$\mu_{n,a} = n - \sum_{i=1}^{n} \frac{n!}{(n-i)!} p_i^{n+a-i}\beta_i, \tag{2.8.2}$$

where $p_i = \phi(\lambda_L i)$ and $\beta_i = G_{i-1}(1 \mid U), U = u_i = \phi(\lambda_L(i+1)) = p_{i+1},\ i = 0, 1, \ldots.$

Thus, $p_i = \exp(-i\lambda_L T_I)$ is the probability that a set of i susceptible individuals who are exposed to a single infective in the same group all escape infection [9].

Alternatively, the mean final size is given in [11] as,

$$\mu_{n,a} = n + a - \sum_{k=0}^{n} \binom{n}{k} \beta_k \phi(\lambda_L k)^{n+a-k}, \tag{2.8.3}$$

where $\beta_0, \beta_1, \ldots,$ are obtained recursively for $k = 0, 1, \ldots,$ in [11] as

$$\sum_{i=0}^{k} \binom{k}{i} \beta_i \phi(\lambda_L i)^{k-i} = k. \tag{2.8.4}$$

2.9 Numerical simulations.

In order to illustrate the threshold behaviour of SIR household epidemic model, we conducted 1000 simulations of a household epidemic for different values of the local and global infection rates, (λ_L, λ_G), using a modified version of the simhouses simulation package of Dr Owen Lyne. This is done using [1] household structure [133, 189, 108, 106, 31] $\times$ 50, where the entries represent number of households which size corresponds to its column. For example 133 is the number of households of size 1, 189 is the number of households of size 2, 108 is the number of households of size 3. The population is made of households of sizes 1 to 5 in which the number of households of each size is 50 times that of [1] and thus a population size of 70700. Also, we have assumed Gamma$(2, 2.05)$ infectious period distribution in [1] which has probability density function, $f_{T_I}(t) = c^2 t \exp(-ct)$, $c > 0$, where $c = 2/4.1$ and mean $E(T_I) = 4.1$ [1,11].

Six pairs of parameter values, (λ_L, λ_G) are considered together with their corresponding threshold parameter in order to study the influence of the infection rates on the occurrence of a global epidemic in the simulation runs. Two columns of histograms of the number of individuals infected from the simulations are presented, with the one on the left having fixed global contact rate and varying local infection rates while those on the right hand side have fixed local infection rate and varying global infection rates.

Form the histograms of the number infected we see that the threshold behaviours exhibits the expected theoretical behaviours such that when $R_* > 1$, then global epidemic occurs with probability $1 - p^a$, where $a = 1$ is the initial number of infectives. The bimodal behaviour of the histograms when $R_* > 1$ further clarify the occurrence a global epidemic in such cases. Thus, large epidemic only occurs when $R_* > 1$ in accordance with [9, 11], also given R_*, the precise values of λ_L and λ_G have little effect on either the number of people infected or the

probability of large epidemic occurring.

Thus, the first two histograms at the top correspond to the case in which $R_* < 1$ and therefore global epidemic never occurred, while the remaining histograms are made of few cases in which a global epidemic occur with bimodal behaviours and few cases in which there is no global epidemic.

In order to disallow the nonglobal epidemic from occurring, we employed a minimum cut-off of the number infected between the epidemics using rejection sampling in which if the number infected in the simulation is less than the cut-off then it is rejected and a re-run is made. This is continued until the simulation run is completed. This is further discussed with examples in section 4.4.

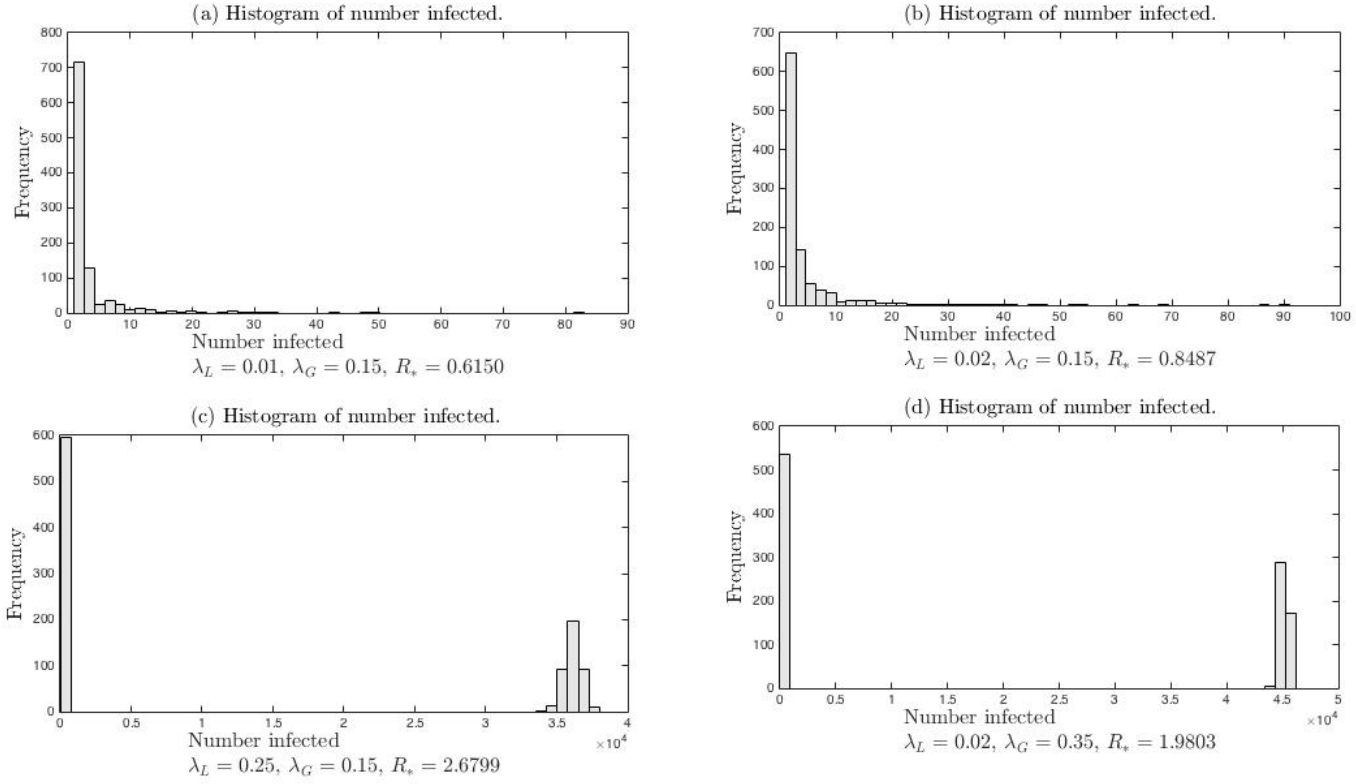

Figure 2.1: Histogram of 1000 simulations of household epidemic with Gamma$(2, 2.05)$ infectious period distribution, parameter estimates from [1] but fifty times its population size and minimum epidemic size of 1.

2.10 Inference on the parameters.

After the model has been chosen, the next stage of the modelling process is estimation of its parameters. Our approach involves constructing likelihood function for the parameters based on assumption of independence of epidemics between households along the lines of [1] independence assumption.

This assumption is contrary to reality and the [9] asumption of dependency of epidemic between households for which our model is based. We have adopted this independence assumption in order to obtain an approximate likelihood function, which can be maximised. We then verify numerically that this provides good estimates using simulation studies as demonstrated in chapters 4.

We have therefore adopted the the maximum likelihood algorithm in [1], using numerical optimization schemes, depending on the dimension of the final size data. Thus, the likelihood function has a single error term for cases where the error terms are the same and two error terms when they are different.

Using the modified version of the simhouses simulation package developed by Dr Owen Lyne which assumes Gamma(2, 2.05) infectious period distribution and the theoretical parameters λ_L and λ_G, we simulate single-type household epidemics without misclassification error, estimate and plot the parameters using the function and subroutines discussed in section 4.2

2.11 Global epidemic.

From section 2.7, the probability of global epidemic for an infection started by a single infectious individual is $1 - q$ [9, 11]. However for an epidemic started by more than one infectious individual, the probability of a global epidemic occurring depends on their configuration [9, 11, 14].

Three cases of initial number of infectives leading to different probability of global epidemic are given by [9]. These include the first case already considered in which an epidemic is started by one infective from a single infectious group, with $n-1$ susceptible, with probability of global epidemic $(1-q)$. In the second case, the epidemic is started by one infectious group containing

i infectives and $n-i$ susceptibles. To obtain the probability of a global epidemic [9] employed the probability generating function of the offspring random variable, defined as $f(s) = E(s^R)$, for $0 \leq s \leq 1$ and conditioning the probability generating function of the offspring distribution on the sum of the infectious periods of the infectives or the severity of the epidemic. We can write the generating function of the number of infected households emanating from a typical infected household, R say, as a mixture of $R_1, R_2, \ldots$, with the respective mixing probabilities $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots$, denoted by $f(s)$ using the method in [6, 9, 11] which is defined as,

$$f(s) = E(s^R) = \sum_{n=1}^{\infty} \tilde{\alpha}_n E(s^{R_n}),$$

where R_n is the total number of global contacts emanating from the household of size n and follows the Poisson distribution with random mean $\lambda_G A_n$, where A_n is the sum of the infectious periods of all the infectives and

$$\begin{aligned} E(s^{R_n}) &= E(E(s^{R_n}|A_n)). \\ &= E(\exp(-\lambda_G A_n(1-s))), \\ &= \phi_{n-1,1}(\lambda_G(1-s)), \end{aligned} \tag{2.11.1}$$

where λ_L and λ_G are the local and global contact rates. $\phi_{n,a} = E(\exp(-\theta A_{n,a}))$ and $A_{n,a}$ is the sum of the infectious periods of the infective individuals in the household epidemic, also called severity of a single household epidemic with initially n susceptibles and a infectives. This is defined in [6, 11] as,

$$\phi_{n,a}(\theta) = \sum_{k=0}^{n} \binom{n}{k} \gamma_k(\theta) \phi(\theta + \lambda_L k)^{n+a-k}. \tag{2.11.2}$$

Here $\gamma_i(\theta)$ for $i = 0, 1, \ldots, n$ are determined recursively by,

$$\sum_{i=0}^{k-1} \gamma_i(\theta) \phi(\theta + \lambda_L i)^{k-i} + \binom{k}{k} \gamma_k(\theta) = 1, (k = 0, 1 \ldots, n).$$

The theoretical properties of the function $\phi_{n,a}$ are discussed in section 3.11 of chapter 3.

2.12 Maximum likelihood estimation.

If $X_{n,j}$ is the number of households of size n with j infectives, (total number of cases), and $P_{n,j}$ is the final size probabilities, (probability of j cases in a household of size n at the end of the epidemic), then each household size, has a separate multinomial distribution for $X_{n,0}, \ldots, X_{n,j}$, $(j = 0, \ldots, n, n = 1, \ldots, max)$, given by [25] as,

$$P(X_{n,0} = x_{n,0}, \ldots, X_{n,j} = x_{n,j}) = \frac{(M_n)!}{\prod_{j=1}^{max}(x_{n,j})!} \prod_{j=0}^{n} P_{n,j}^{x_{n,j}}, \tag{2.12.1}$$

where M_n is the number of household of type n among the infected households.

By assuming independence of epidemics in each household in accordance with [1], the likelihood function is referred to as approximate likelihood function of the parameters λ_L and π, [9] given by,

$$L(\lambda_L, \pi) = \frac{(M_{n.})!}{\prod_{i=1}^{max}(x_{n,j})!} \prod_{i=1}^{max} \prod_{j=0}^{n} P_{n,j}(\lambda_L, \pi)^{x_{n,j}}, \tag{2.12.2}$$

where $P_{n,j}$ are the final size probabilities, n is the household size, π is the probability of avoiding infection from outside the household, λ_L is the local contact rate, $x_{n,j}$ is the final size data defined as the number of households of size n with j number of infectives, max is the maximum household size, and M_n is the number of households of size n among the infected households.

The approximate likelihood function for cases when the final size epidemic data is subject to misclassification will be discussed in chapter 4.

Using logarithm in equation (2.12.2) for ease of computation and simplification, we can express the approximate likelihood function in terms of its loglikelihood as,

$$l(\lambda_L, \pi) = \log(M_{n.})! - \sum_{i=1}^{max} \log(x_{n,j})! + \sum_{i=1}^{max} \sum_{j=0}^{n} x_{n,j} \log P_{n,j}. \tag{2.12.3}$$

The approximate loglikelihood function of the theoretical parameters, λ_L and π can then

be computed using appropriate numerical optimization along the lines of the computational algorithm given in [1].

We have developed Matlab programs using the Nelder-Mead fminsearch simplex numerical algorithm referred to here as two dimensional numerical optimization to estimate the parameters.

Chapter 3

Theoretical properties of the parameters of the stochastic SIR household epidemic model.

3.1 Introduction.

In this chapter, we studied the theoretical properties of the parameters and functions of the stochastic SIR household epidemic model beginning with the mean final size of the household epidemic and the beta function for small and large local infection rates. In section 3.7, we discussed the properties of the threshold parameter. In section 3.8, we examined the proportion of the initial susceptibles infected in a household epidemic, while in section 3.11, we also examined the Gamma function for the generating function of the number of infected households from a typical infected household.

These terms are fully explained in the indicated sections of this chapter.

3.2 The mean final size of single household epidemic.

The mean final size of a single household epidemic is given in [9] and is defined as the average number of initial susceptibles that are ultimately infected, including the initial number of

infectives, at the end of the disease outbreak expressed as

$$\mu_{n,a} = n + a - \sum_{k=0}^{n} \binom{n}{k} \beta_k \phi(\lambda_L k)^{n+a-k},$$

where n is the total number of susceptibles, a is the initial number of infectives at the beginning of disease outbreak, β_k are functions of λ_L and the infectious period distribution, obtained for $k \in \mathbb{Z}_+$ from the triangular equation in [6] as,

$$\sum_{i=0}^{k} \binom{k}{i} \beta_i \phi(\lambda_L i)^{k-i} = k, \quad k = 0, 1, 2, \ldots,$$

where, $\phi(\theta) = E(\exp(-\theta T_I))$, is the moment generating function of the infective period, T_I, and λ_L is the local contact rate. This can be expanded as,

$$\binom{k}{0} \beta_0 \phi(\lambda_L.0)^{k-0} + \binom{k}{1} \beta_1 \phi(\lambda_L.1)^{k-1} + \cdots + \binom{k}{k-1} \beta_{k-1} \phi(\lambda_L \cdot (k-1))^1 + \beta_k \phi(\lambda_L.k)^0 = k.$$

Observe that if $k = 0$, then $\beta_0 = 0$. Thus we can ignore the first term and express the equation as,

$$\binom{k}{1} \beta_1 \phi(\lambda_L.1)^{k-1} + \binom{k}{2} \beta_2 \phi(\lambda_L.2)^{k-2} + \cdots + \binom{k}{k-1} \beta_{k-1} \phi(\lambda_L \cdot (k-1))^1 + \beta_k = k.$$

We can also rearrange it as,

$$\beta_k = k - \sum_{i=1}^{k-1} \binom{k}{i} \beta_i \phi(\lambda_L i)^{k-i}. \tag{3.2.1}$$

3.3 Properties of β_k for small and large local infection rates.

If $\lambda_L \to 0$, then $\phi(\lambda_L) = E(\exp(-\lambda_L T_I)) \to 1$, $\forall\ T_I$ and equation 3.2.1 reduces to

$$\sum_{i=1}^{k} \binom{k}{i} \beta_i = k.$$

It follows that, if $\lambda_L \to 0$, β_k can be expressed as,

$$\beta_k = k - \sum_{i=1}^{k-1} \binom{k}{i} \beta_i.$$

Theorem 2. *If $\lambda_L = 0$, then $\beta_k = 0$, $\forall k \in \mathbb{Z}_+ - \{1\}$ and $\beta_1 = 1$ when $k = 1$.*

Proof. Using mathematical induction, we will show that $\beta_k = 0$, $\forall k \in \mathbb{Z}_+ - \{1\}$, whenever $\lambda_L = 0$.

From the arguments in equation (3.2.1), we know that $\beta_0 = 0$, when $k = 0$, also when $k = 1$, $\beta_1 = 1$, however when $k = 2$, then $\beta_2 = 0$.

Using mathematical induction, we want to show that $\beta_k = 0$, $\forall k \in \mathbf{Z}_+ - \{1\}$.

We assume the induction hypothesis holds for $\forall n \in \{2, \ldots, k\}$ and show that it also holds for β_{k+1}.

For $k + 1$, we have,

$$\beta_{k+1} = k + 1 - \sum_{i=1}^{k} \binom{k+1}{i} \beta_i.$$

Using $\binom{k+1}{m} = \binom{k}{m} + \binom{k}{m-1}$, $\forall m, k \in \mathbb{Z}_+$, reduces the problem to the form,

$$\beta_{k+1} = k + 1 - \sum_{i=1}^{k-1} \binom{k}{i} \beta_i - \beta_k - \sum_{i=1}^{k} \binom{k}{i-1} \beta_i.$$

Replacing the term, $\sum_{i=1}^{k-1} \binom{k}{i} \beta_i$ with $k - \beta_k$ and simplifying gives,

$$\beta_{k+1} = 1 - \sum_{i=1}^{k} \binom{k}{i-1} \beta_i.$$

For example, when $k = 2$ we get, $\beta_3 = 1 - \binom{2}{0}\beta_1 + \binom{2}{1}\beta_2$. Substituting $\beta_0 = 1$ and $\beta_2 = 0$

gives the required results.

Since by hypothesis, $\beta_i = 0$, for $i = 0, 2, 3 \ldots k$, and $\beta_1 = 1$, the result follows that $\beta_{k+1} = 1 - \beta_1 = 0$. Therefore, the result follows by induction.

□

If $\lambda_L \to \infty$, then $\phi(\lambda_L) = E(\exp(-\lambda_L T_I) \to 0$, for all positive random variables T_I and the expression $\sum_{i=1}^{k-1} \binom{k}{i} \beta_i \phi(\lambda_L i)^{k-i}$, gives $\beta_k = k$ when $\lambda_L \to \infty$.

In figures 3.1 (a) and (b), we have plotted the β_k, as a function of λ_L while holding other parameters as $n = 6$, Gamma(a)=2, Gamma(b)=2.05 and $c = 1$ number of initial infective, for two extreme values of λ_L, that is when $\lambda_L \to \infty$ and when $\lambda_L \to 0$. We have adopted the Gamma infectious period distribution to enables us compare our results with those of [1] who also employed the Gamma(2, 2.05) infectious period distribution.

The behaviour of β_k is found to be consistent with our theoretical studies. When λ_L becomes very large, β_k becomes asymptotic to k, while as λ_L approaches 0, so also is β_k. This can be seen from figures 3.1.

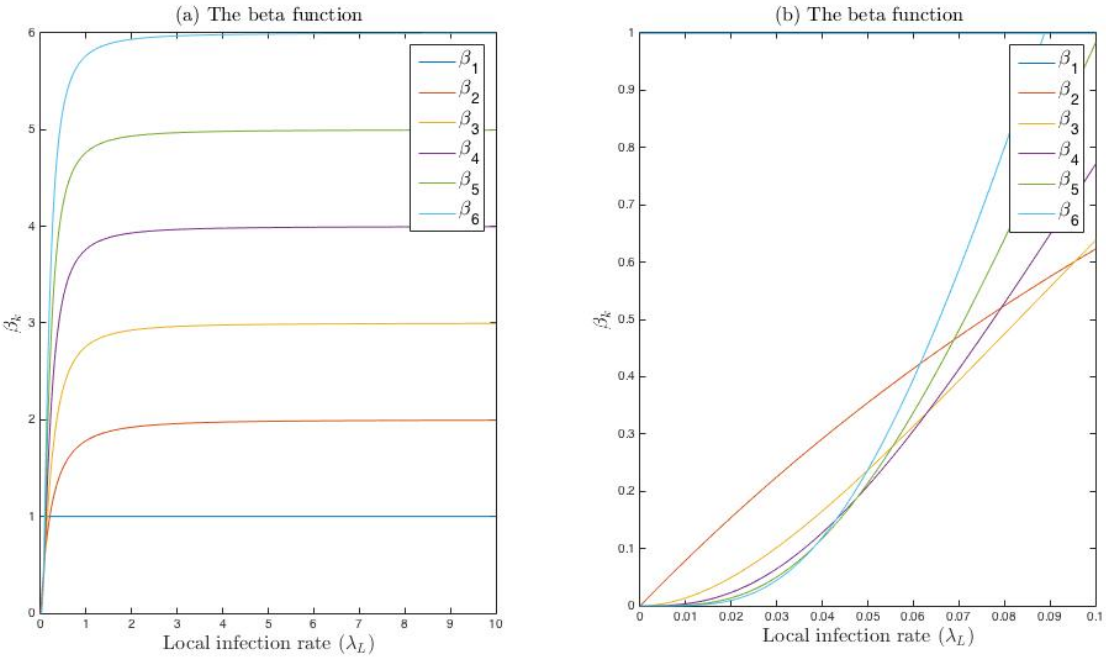

Figure 3.1: The beta function with increasing λ_L.

In figure 3.1, we plotted the beta function as a function of λ_L, using Gamma(a, b) infectious period distribution with parameters Gamma$(a) = 2$, Gamma$(b) = 2.05$. We see that with increasing λ_L, the function β_k also increases and tends to $k = 1, \ldots, 5$, where $\beta_0 = 0$, while as λ_L tends to zero, β_k also tends to zero except β_1 which assumes the value 1.

3.4 The mean final size of the single household epidemic for small λ_L.

Using the properties of β_k and since $\phi(\lambda_L) \to 1$, if $\lambda_L \to 0$, the expression for the mean final size reduces to

$$\mu_{n,a} = n + a - \sum_{k=0}^{n} \binom{n}{k} \beta_k,$$

where $n+a$ is the household size, n and a are the number of initial susceptibles and infectives.

Since $\beta_k \to 0\ \forall k \in \mathbb{Z}_+ - \{1\}$ with $\beta_1 = 1$, when $\lambda_L \to 0$, the expression for the mean final size reduces to,

$$\mu_{n,a} = n + a - \binom{n}{0}\beta_0 + \binom{n}{1}\beta_1.$$

Putting the values of $\beta_0 = 0$ and $\beta_1 = 1$ into the expression yields the value of the mean final size of a single household epidemic, when $\lambda_L \to 0$,

$$\mu_{n,a} = n + a - n = a.$$

This means that if there are no local contacts between susceptible and infective individuals in the household, there will be no new infections and the ultimate number of infected individuals at the end of the epidemic will be the initial number of infectives.

3.5 The mean final size of the single household epidemic, for large local infection rates.

If $\lambda_L \to \infty$, then $\phi(\lambda_L) = E(\exp(-\lambda_L T_I)) \to 0$, since T_I is a non-negative random variable and since β_k assumes the values $k \in \mathbb{Z}_+$, we can write the mean final size equation as,

$$\mu_{n,a} = n + a - \left(\binom{n}{0}\beta_0\phi(\lambda_L.0)^{n+a-0} + \binom{n}{1}\beta_1\phi(\lambda_L.1)^{n+a-1} + \cdots + \binom{n}{k}\beta_k\phi(\lambda_L.k)^{n+k-1}\right).$$

We know that if $\lambda_L \to \infty$, then $\phi(\lambda_L) \to 0$. The question then is, can $n + a - k$ be zero, since if $n + a - k$ is zero then the expression $\phi(\lambda_L.k)^{n+a-k}$ reduces to 1. Since k is only defined for $k = 0, 1, 2, \ldots, n$ and a is not zero, if a is zero then there will be no infection in the household and so no susceptible individuals will be subjected to any infection pressure and so $k < n + a$, $\forall a \in \mathbb{Z}_+ - \{0\}$,

However, if $k = 0$ then $\beta_k\phi(\lambda_L.k)^{a+n-k}$ reduces to zero, since $\beta_0 = 0$.

If $a \neq 0$, $k \neq 0$, then $n + a\ > k$.

Under this assumption, $\beta_k \phi(\lambda_L.k)^{a+n-k} \to 0$ and the summation terms on the right hand of mean final size will collapse to zero with the mean final size given by the remaining term as,

$$\mu_{n,a} = n + a.$$

This means that everybody will be infected at the end of the epidemic outbreak, which is possible for highly infectious diseases with large local contact rate. The role of these parameters on household disease transmission is crucial and any effective intervention, and control strategies must take this into consideration.

3.6 Further properties of the mean final size.

The influence of the local contact rate and other parameters of the mean final size on its behaviour is further studied using graphs by varying some of the parameters while holding others constant.

For example λ_L is considered as an independent variable and plotted with the mean final size over the range of values [0, 1] as in figure 3.2, with $n = 6$, Gamma(a) =2, Gamma(b)=2.05, $c = 1$.

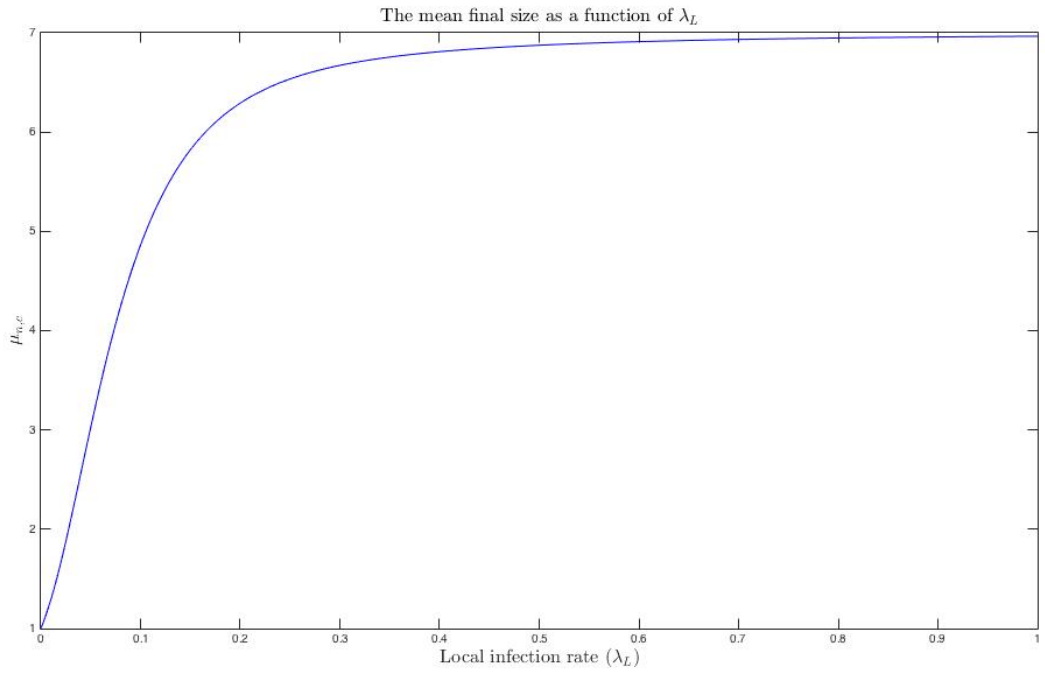

Figure 3.2: The mean final size as function of the local infection rate.

In figure 3.2, the mean final size of the household epidemic with $n = 6$ initial susceptibles and $c = 1$ initial infectives $\mu_{n,c}$, is plotted as a function of λ_L by assuming Gamma(a, b) as the infectious period distribution with the parameters, Gamma$(a) = 2$, Gamma$(b) = 2.05$. We see that the mean final size of the household epidemic increases with increasing λ_L. The mean final size therefore reduces to the initial number infected when it is zero in line with the discussions in section 3.4.

The mean final size increases rapidly towards the maximum household size in response to continuous increase in the value of λ_L, as shown, which is in agreement with its theoretical properties.

If $\lambda_L \to 0$, the mean final size $\mu_{n,a} \to a$. This shows that the magnitude of local contact rate for within household infection contributes to the level of disease transmission.

We also examined the behaviour of the mean final size given initial number of infectives and susceptibles for varying local contact rate for the following cases, $\lambda_L = 0.001, 0.05, 1$ with varying initial number of infectives, c and initial number of susceptible $n = 2$, while $\lambda_L = 0.001, 0.1, 1$ with varying initial number of susceptibles, n and initial number, of infective $c = 1$ respectively. Identical behaviour as λ_L becomes large is observed. This can be seen in figures 3.3 and 3.4.

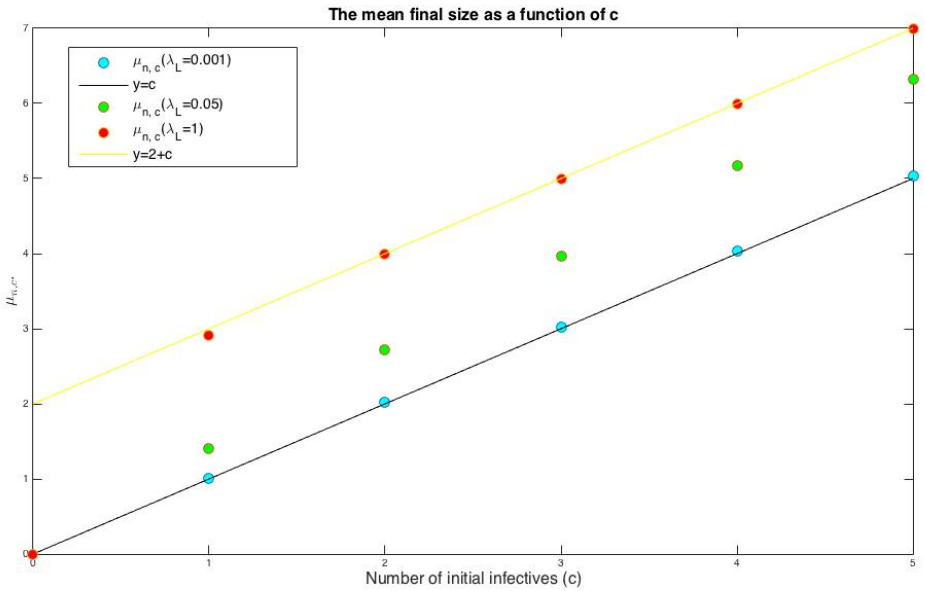

Figure 3.3: The Mean final Size as function of the number of initial infectives

In figure 3.3, the mean final size is plotted as a function of the number of initial infectives in a single household epidemic and varying local infection rate $\lambda_L = 0.001, 0.05, 1$ and Gamma(a, b) infectious period distribution, having parameters, Gamma(a)=2, Gamma(b)=2.05. As λ_L becomes sufficiently large. the mean final size increases and becomes asymptotic to the line, $y = 2 + c$, which forms its upper bound, with 2 as the number of initial susceptibles, where c the initial number of infectives.

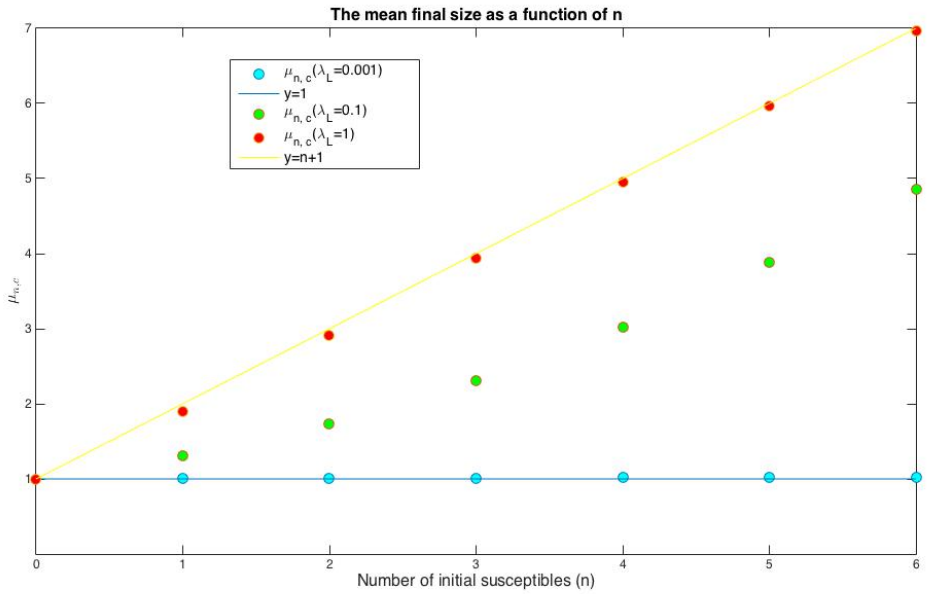

Figure 3.4: The Mean Final Size as function of number of the initial susceptibles.

In figure 3.4, the mean final size is plotted as a function of the initial number of susceptibles, n and varying values of $\lambda_L = 0.001, 0.1, 1$, Gamma(a,b) infectives period distribution with parameters, Gamma(a)=2, Gamma(b)=2.05, $c = 1$. It is found that the mean final size approaches the line $y = n + 1$, as $\lambda_L \to \infty$, where 1 is the initial number of infectives. The line $y = n + 1$ and $y = 1$ are its upper and lower bounds.

The mean final size tends to approach the line $y = 2 + c$, for large values of λ_L, where 2 is the number of susceptibles in the household, which is its upper bound, as further increase in λ_L makes no contribution to the mean final size. In the case of n, which is the number of initial susceptibles, the mean final size becomes asymptotic to the line $y = n + 1$, forming its upper bound. Where 1 is the number of initial infectives in the household. This shows that as $\lambda_L \to \infty$, these lines are representative of the mean of the final size. These behaviours are shown in figures 3.3 and 3.4 respectively.

3.7 Properties of the threshold parameter for small and large local infection rates.

The threshold parameter as defined in section 2.7 is a function of both the local and global infection rates. If the global infection rate, $\lambda_G \to 0$, then the threshold parameter will be zero, on the contrary if $\lambda_L \to 0$, then β_k will all be zero except $\beta_1 = 1$ in accordance with the properties of β_k and the resulting mean final size $\mu_{n-1,1}$ of the household with $n - 1$ initial susceptibles and 1 initial infective will be the initial infective, which under this definition is $\mu_{n-1,1} = 1$ with the threshold parameter given by

$$R_* = \lambda_G E(T_I) \sum_{n=1}^{\infty} \tilde{\alpha}_n,$$

Since $\tilde{\alpha}_n$ are probabilities, their summation will be 1, reducing the threshold parameter to

$$\begin{aligned} R_* &= \lambda_G E(T_I), \\ &= R_0. \end{aligned}$$

The household threshold parameter R_* is expressed in terms of R_0 in [6, 9, 11] as,

$$R_* = R_0 \mu,$$

where $R_0 = \lambda_G E(T_I)$ and $\mu = \sum_{n=1}^{\infty} \tilde{\alpha}_n \mu_n$ is the mean amplification factor owing to internal spread within the household as in section 2.7. Where R_0 defined in section 2.7, is the basic reproductive ratio for homogeneous mixing population, in which everyone is assumed to have similar characteristics without consideration for heterogeneity in infectivity and susceptibility. It is a threshold parameter for a population in which the household size is one. It can loosely be defined as the average number of infectives generated by a single infected individual in a completely susceptible population throughout its infectious period.

The behaviour of the threshold parameter for varying local infection is studied for some global infection rates, $\lambda_G = 0.01, 0.02, 0.03, 0.04$, and $\lambda_G = 0.1, 0.2, 0.3, 0.4$ respectively, Gamma(a, b) infectious period distribution with parameters, $a = 2, b = 2.05$ and assuming single initial infective, $c = 1$, in the household. We found in each of the cases that large global infection rate leads to corresponding large threshold parameter. Thus the threshold parameter is linearly influenced by the level of the global contact rate.

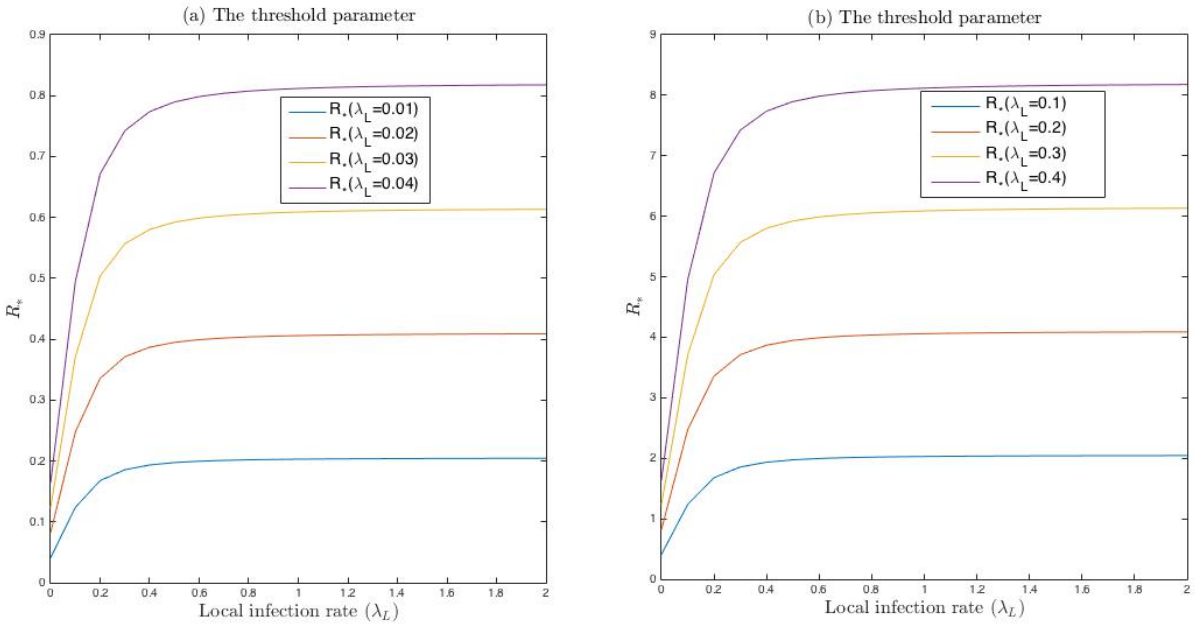

Figure 3.5: The threshold parameter with varying local infection rate.

In figure 3.5, we have plotted the threshold parameter for varying local infection rate defined in the region $\{\lambda_L : 0 \leq \lambda_L \leq 2\}$, with stepsize of 0.05, for the following global infection rates $\lambda_G = 0.01, 0.02, 0.03, 0.04$, and $\lambda_G = 0.1, 0.2, 0.3, 0.4$ respectively and Gamma$(2, 2.05)$ infectious period distribution, and one initial infective, $c = 1$.

3.8 Proportion of the initial susceptibles that are ultimately infected.

The proportion of the initial susceptible individuals that are ultimately infected by the epidemic, denoted by z, is given in [11] as

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n n^{-1} \sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} \mu_{n-k,k}. \tag{3.8.1}$$

Equation (3.8.1) is the weighted average of the number of infectives in a single household epidemic with Binomial distributed number of infectives k, and the remaining $n-k$ susceptibles avoid infection from outside the household of size n.

In equation (3.8.1), $\tilde{\alpha}_n$ is the probability that a randomly selected individual resides in a household of size n, π is the probability that a given individual avoids global infection, which is approximately given in [9, 11] as,

$$\pi = \exp(-\frac{\lambda_G}{N} z N E(T_I)) = \exp(-\lambda_G z E(T_I)). \tag{3.8.2}$$

Where $NzE(T_I)$ is the total person units of infection present throughout the epidemic, N is the total number of individuals in the household and z is the proportion of the initial susceptibles ultimately infected.

Suppose global epidemic has occurred with the proportion of individuals ultimately infected, $z \in [0, 1]$, then equations (3.8.1) together with (3.8.2) gives an implicit equation for z. Here $z = 0$ is always a solution and the only solution if $R_* \leq 1$. A second solution in $0 < z < 1$ exists only if $R_* > 1$.

This is better understood by expressing equation (3.8.2) in the form $y = z = g(z)$ where $y = z$, $y = g(z)$. Here $g(z)$ is the right hand side of equation (3.8.2) and the unique solution of the equation is found at the point of intersection of $y = z$ and $y = g(z)$ nearest to the origin for which $R_* > 1$. Now let the generating function of the offspring random variable R be defined as $E(z^R) = g(z)$ and P_k be its distribution. Then $g(z) = \sum_{k=0}^{\infty} P_k z^k$ with $g'(1) = \sum_{k=1}^{\infty} k P_k$ which is equal to R_*.

For example, using numerical calculation with [1] final size epidemic data and a range of values of $\pi = 0.2, 0.4, 0.6, 0.9$ and $\lambda_L \in [0, 1]$, we found that as π increases towards its upper boundary, the unique root of $z = g(z)$ decay as theoretically expected. Thus z depends on the magnitude of π such that the more the susceptibles in the households avoid global infection, the less the proportion ultimately infected at the end of the epidemic as demonstrated in figure 3.6.

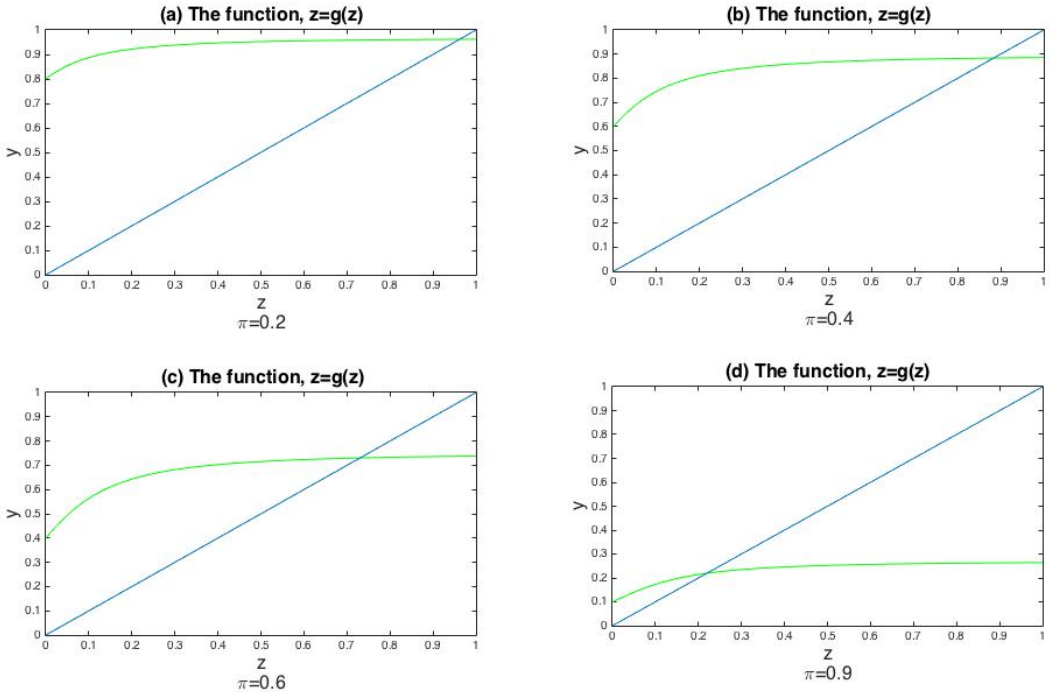

Figure 3.6: The proportion of the initial susceptible ultimately infected at the end of the epidemic in the presence of varying π.

3.9 Proportion of the initial susceptibles that are ultimately infected at the lower boundary of the local infection rate.

If the local contact rate $\lambda_L \to 0$, then the mean final size of a household with k initial infectives and $n-k$ initial susceptibles is, $\mu_{n-k,k}(0) = k$. We can express z as,

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n n^{-1} \sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} k. \tag{3.9.1}$$

Since,

$$E(K) = \sum_{k=0}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} k = n(1-\pi),$$

where $\binom{n}{k}(1-\pi)^k \pi^{n-k}$ is the probability that k susceptibles individuals are infected with probability $(1-\pi)^k$, while the remaining $n-k$ escape infection with probability π^{n-k}.

The number of infectives k, in the household is distributed as a binomial random variable, with parameters, n and $(1-\pi)$. Here $E(K)$ is the mean number of infected susceptibles in the household. Substitution of the mean number of susceptible individuals infected, $E(K) = n(1-\pi)$ into the expression for z gives,

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n (1-\pi) = (1-\pi).$$

This can further be simplified as,

$$z = 1 - \pi = 1 - \exp(-\lambda_G z E(T_I)). \tag{3.9.2}$$

This is the governing equation of z for the single population S-I-R deterministic epidemic model.

3.10 Proportion of the initial susceptibles that are ultimately infected near the upper boundary of the local infection rate.

If $\lambda_L \to \infty$, then the mean final size in equation, $\mu_{n-k,k}$ for $k > 0$ reduces to n and the expression for z becomes,

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n n^{-1} \sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} n. \tag{3.10.1}$$

Since,

$$\sum_{k=0}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} = 1,$$

we will have,

$$\sum_{k=0}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} = \pi^n + \sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k},$$

where $p(K=0) = \pi^n$ is the probability that every susceptible in a household of size n avoids global inflection. We can write

$$\sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} = 1 - \pi^n.$$

We can then express z as,

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n n^{-1} \sum_{k=1}^{n} \binom{n}{k} (1-\pi)^k \pi^{n-k} n = \sum_{n=1}^{\infty} \tilde{\alpha}_n (1-\pi^n),$$

$$z = \sum_{n=1}^{\infty} \tilde{\alpha}_n (1 - \exp(-n\lambda_G z E(T_I))),$$

where $\pi^n = \exp(-n\lambda_G z E(T_I))$. Further simplification of z gives,

$$z = 1 - \sum_{n=1}^{\infty} \tilde{\alpha}_n \exp(-n\lambda_G z E(T_I)). \tag{3.10.2}$$

3.11 Theoretical properties of the Gamma function and global epidemic.

The expressions for the gamma function from the triangular equation for the generation function of the number of infected households from a typical infected household R called the offspring distribution discussed in section 2.11 is further explored. Recall that this generating function is given by

$$\phi_{n,a}(\theta) = \sum_{k=0}^{n} \binom{n}{k} \gamma_k(\theta) \phi(\theta + \lambda_L.k)^{n+a-k}. \tag{3.11.1}$$

Where $\gamma_i(\theta)$ for $i = 0, 1, \ldots, n$ are determined recursively by,

$$\sum_{i=0}^{k-1} \gamma_i(\theta) \phi(\theta + \lambda_L.i)^{k-i} + \binom{k}{k} \gamma_k(\theta) = 1, \ (k = 0, 1 \ldots n). \tag{3.11.2}$$

The expression for the gamma function in equation (3.11.2) can be simplified for every $k = 0, 1, \ldots n$ as follows,

$k = 0$, gives,

$$\binom{0}{0} \gamma_0(\theta) \phi(\theta + \lambda_L.0)^{0-0} = 1,$$

$$\gamma_0(\theta) = 1, \forall \theta \geq 0. \tag{3.11.3}$$

Thus, we can write gamma $\gamma_k(\theta)$ as,

$$\gamma_k(\theta) = 1 - \sum_{i=0}^{k-1} \binom{k}{i} \gamma_i(\theta) \phi(\theta + \lambda_L i)^{k-i}, \ k = 1, 2 \ldots n, \ \forall \theta \geq 0, \tag{3.11.4}$$

where $\phi(\theta) = \exp(-\theta T_I)$, T_I is the infectious period of an infected individual, whose choice is arbitrary but must be specified with known moment generating function. We have assumed T_I to follow the Gamma(a, b) distribution as in [1].

If λ_L and θ approach zero simultaneously then we can derive an expression for $\gamma(\theta)$, for

$\theta = 0$. Since under this assumption $\phi(0)$ goes to 1, and we will have,

$$\gamma_k(0) = 1 - \sum_{i=0}^{k-1} \binom{k}{i} \gamma_i(0),\ k = 1, 2 \ldots n,$$

Theorem 3. *If* $\lambda_L = 0$, *and* $\theta = 0$, *then* $\gamma_k(0) = 0\ \forall k \in \mathbf{Z}_+$.

Proof. We prove by induction that $\gamma_k(0) = 0\ \forall k \in \mathbf{Z}_+$.

When $k = 1$, equation (3.11.5), reduces to, $\gamma_1(0) = 1 - \binom{1}{0}\gamma_0(0) = 0$.

When $k = 2$, $\gamma_2(0)$ reduces to,

$$\gamma_2(0) = 1 - \left(\binom{2}{0} \gamma_0(0) + \binom{2}{1} \gamma_1(0) \right),$$

using $\gamma_0(0) = 1$ and $\gamma_1(0) = 0$, we get $\gamma_2(0) = 1 - 1 = 0$.

We assume, this expression holds for any $k \in \mathbf{Z}_+$ and show that it holds for $k + 1$.

For $k + 1$, we will have,

$$\gamma_{k+1}(0) = 1 - \sum_{i=0}^{k} \binom{k+1}{i} \gamma_i(0).$$

Since, we can express $\binom{k+1}{i}$ as $\binom{k}{i} + \binom{k}{i-1}$, we can write

$$\gamma_{k+1}(0) = -\sum_{i=1}^{k} \binom{k}{i-1} \gamma_i(0).$$

Thus, $\gamma_{k+1}(0) = 0$.

The hypothesis, holds for $k + 1$, and in general, $\gamma_k(0) = 0, \forall k \in \mathbf{Z}_+$. □

If $\theta \to \infty$, then it is obvious that $\gamma_k(\theta) = 1$ since $\phi(\theta + \lambda_L.i)^{k-i} \to 0$, where $k - i \geq 1$. Similarly, $\phi(\theta + \lambda_L.k)^{n+a-k}$ will be zero, since $n + a - k \geq a$ and $a \geq 1$. Hence, the generating function reduces to 0. It follows that $f(s) = 0$, and $s = 0$. There will a global epidemic with probability 1.

Using the representation of $\theta = \lambda_G(1 - s)$ in these studies, in line with [6, 9] and [11], we see that $\gamma_k(\theta)$ is a function of both λ_L and λ_G respectively. If $\theta \to 0$, then either $\lambda_G \to 0$, for

$s \in (0, 1]$, or $s = 0$, for $\lambda_G > 0$.

If $\lambda_G \to 0$, for some $s \in [0, 1)$, and $\lambda_L \to 0$, then $\gamma_k(0) = 0$, for $\forall k \in \mathbf{Z}_+ - \{0\}$, $\phi_{n,a}(0) = 1$, $f(s) = 1$. There will be nonglobal epidemic with probability 1. The probability of a global epidemic is 0.

Chapter 4

Fitting the SIR household model to final size epidemic data.

4.1 Introduction.

Having specified the model and explored the behaviour of its parameters, it is then necessary to fit it to the final size epidemic data in section 1.9, using maximum likelihood estimation along the lines of [1] for which the epidemic in each household is assumed independent of the epidemic in other households.

This chapter is concerned with fitting the stochastic SIR household epidemic model to two dimensional final size data made of true infectives in the households.

Using simulation studies, we present plots of the estimates and table of mean, standard deviation and the root mean square to give further insights into their precision.

These are accomplished using fifty times population size in [1] and minimum epidemic size of 1000.

4.2 Model fitting to the two dimensional final size data.

Sometimes individuals are observed correctly as true positives and negatives in the households. The final size data is then made of the number of households with true outbreak sizes, as

discussed in section 1.9. Fitting the two dimensional model to such final size data employs [1] maximum likelihood algorithm discussed in section 2.12.

Using simulation studies we implement the estimation with the function csimhouses.m as follows.

1. Run the function csimhouses to simulate two dimensional household epidemic data, denoted here as mat with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, and λ_G, minimum epidemic size and number of repetitions required. The parameters are then estimated and plotted. Their mean, standard deviation, root mean square error are also computed and using the following subroutines.

a.) fminsearch2$(n, a, b, \max)$ which maximizes the loglikelihood function using starting value according to [24].

b.) LampaiD(mat), provides starting values for the estimates according to [24].

b.) negloglik2(y, n, a, b, mat) computes the negative loglikelihood using starting values of the parameters according to [24], the parameters of Gamma(a, b) infectious period distribution and the final size epidemic data.

c.) final_sizep$(a, b, \pi, n, \lambda_L)$ calculates the final size probabilities required by the subroutine negloglik2(y, n, a, b, mat) from π, λ_L, and Gamma(a, b) infectious period distribution.

e.) pinf2$(a, b, \pi, \lambda_L \text{houses})$, calculates z and λ_G, from π, λ_L, maximum household size n and parameters of Gamma(a, b) infectious period distribution.

f.) RSTER2$(a, b, c, \lambda_L, \lambda_G, houses)$ calculates the threshold parameter, R_* from the theoretical parameters, λ_L, λ_G and the parameters of Gamma(a, b) infectious period distribution.

The likelihood function in equation (2.12.2) is referred to as the approximate likelihood because of the assumption of independence of the epidemics in each household in [1] which is not consistent with the assumption in [9]. The assumption is not true but it is adopted to allow the use of the maximum likelihood algorithm in line with [1] for the estimation of the parameters.

The process is such that the starting values for π and λ_L are obtained according to [24] from equations (4.2.1) and (4.2.2).

For example estimating π, requires equation (4.2.1) to be used in evaluating the starting

value given as

$$\hat{\pi} = {}^{1}\!/\!_{n} \sum_{s=1}^{\max} n_s \left(\frac{n_{0,s}}{n_s} \right)^{{}^{1}\!/\!_{s}}, \tag{4.2.1}$$

where n is the total number of households, max is the maximum household size, n_s is the number of households of sizes s and $n_{j,s}$ is the number of households of size s in which the size of the outbreak is $j = 0, 1, \ldots, s$. i.e. number infectives in the household of size s. Observe that $\sum_{j=0}^{s} n_{j,s} = n_s$ and $\sum_{s=1}^{\max} n_s = n$ respectively.

Here, ${}^{n_{0,s}}\!/\!_{n_s}$ is an unbiased estimate of $P_0(s) = \pi^s$, where $P_0(s)$ is the probability of zero infectives in the household of size s, which can also be read as the probability that all the susceptibles in the household of size s avoid global infection.

Then $({}^{n_{0,s}}\!/\!_{n_s})^{1/s}$ provides estimates of π for the household sizes $s = 1, 2, \ldots, max$. Pooling the estimates together [24] gave the initial estimates in equation (4.2.1).

For the local infection rate, a reasonable estimate for λ_L for the household size s is given by [24] as, $({}^{n_{1,s}}\!/\!_{(n_s - n_{0,s})})^{1/(s-1)}$ and is unity when $n_{0,s} = n_s$.

Pooling the estimates together as in [24], the estimate of λ_L is started using,

$$\hat{\lambda}_L = \frac{1}{\sum_{s=2}^{\max}(n_s - n_{0,s})} \sum_{s=2}^{\max} (n_s - n_{0,s})\left(\frac{n_{1,s}}{n_s - n_{0,s}}\right). \tag{4.2.2}$$

Consider an alternative estimation techniques for the theoeretical parameters. For example if we know the pair of parameters, (λ_L, λ_G), then by defining a new functional denoted by D, which is the sum of square difference between the old and new values of π and between the old and new values of z defined as,

$$D = (\pi_{old} - \pi_{new})^2 + (z_{old} - z_{new})^2,$$

$$\pi_{New} = \exp(-\hat{\lambda}_G z_{old} E(T_I),$$

$$z_{New} = \sum_{n=1}^{\infty} \tilde{\alpha_n} n^{-1} \sum_{k=1}^{n} \binom{n}{k} (1 - \pi_{old})^k \pi_{old}^{n-k} \mu_{n-k,k} \hat{\lambda}_L.$$

We can then adopt the Nelder-Mead fminsearch simplex numerical algorithm on D to find the values of z and π. The parameters are estimated as follows.

2. Run the function zpfun$(x, a, b, \lambda_L, \lambda_G, \text{houses})$, which uses fminsearch algorithm to calculate π and z from the theoretical parameters, λ_L, λ_G, vector of the household sizes and the parameters of Gamma(a, b) infectious period distribution. It uses the following subroutines.

a.) zp_old$(x, a, b, \lambda_L, \lambda_G, \text{houses})$, calculates the the function D, which is the sum of the square difference between the old and new value of z and between the old and new value of π and maximized by fminsearch simplex algorithm using vector of starting values for the parameters.

b.) RSTER2$(a, b, c, \lambda_L, \lambda_G, houses)$ calculates the threshold parameter R_* from the theoretical parameters, λ_L, λ_G, and those of Gamma(a, b) infectious period distribution and the initial number infected c.

4.3 Replication of published results.

It is necessary to examine the performance of our program functions by assessing and comparing the parameter estimates from them with those of published results in [1, 9].

If they are the same, then it will mean that our program functions are working well and can be employed to fit the stochastic SIR household epidemic model to two dimensional household final size epidemic data.

We do this by fitting the stochastic SIR epidemic model to [1] final size epidemic data in table 1.2 using the first method in section 4.2 and the same household structure and size in [1], and assuming Gamma(a, b) infectious period distribution as in [1, 9], where $a = 2$, $b = 2.05$, and density function $f(t) = c^2 t \exp(-ct)$, $t > 0$, $c = {}^2/_{4.1}$, single initial infective, we then estimated the parameters, λ_L and π in [1, 9].

We know that the parameters are estimated by [1] as $\lambda_L = 0.0446$ and $\pi = 0.8674$ with population size of 1414 wih maximum household size $n = 5$. While [9] estimated $\lambda_G = 0.1955$, $z = 0.1775$ and $R_* = 1.1303$ with the same population size and assuming Gamma$(2, 2.05)$ infectious period distribution. Using the first method in section 4.2 , we replicated the estimates of λ_L and π as follow,

Parameter	Published results.		Calculated results from the codes.
	[1]	[9]	
$\hat{\lambda_L}$	0.0446	0.0446	0.0446
$\hat{\lambda_G}$	N/A	0.1955	0.1955
$\hat{\pi}$	0.8674	0.8674	0.8674
$\hat{z}$	N/A	0.1775	0.1775
$\hat{R_*}$	N/A	1.1303	1.1304

Table 4.1: We compared estimates from published results in [1,9] with those from our Matlab programs discussed in section 4.2. The notation N/A means estimate of the parameter not provided by the author.

The observed proportion infected is computed as,

$$\frac{1(23+27+23+20+9)+2(13+6+16+5)+3(7+8+2)+4(2+1)+5(1)}{1(133)+2(189)+3(108)+4(106)+5(31)}$$
$$= \frac{250}{1414} = 0.1768.$$

We have seen that our program functions give estimates which are the same in the numerical accuracy used to those in [1, 9] respectively. Our program functions are working well and can therefore be used to fit the stochastic SIR household epidemic model to the two dimensional household final size epidemic data.

4.4 Simulation and inference.

We have adopted the likelihood function for the non misclassified final size data in equation (2.12.2), which we have referred to as approximate likelihood function as discussed in section 4.2.

Using the assumption of independence of epidemic between households in [1] and since each household size (number of cases) has separate multinomial distribution given in equation (2.12.1), we can express the approximate likelihood function as in equation (2.12.2).

The parameters of the approximate likelihood function which are the local infection rate and the probability of avoiding infection from outside the households, λ_L and π are then

estimated using the program function and subroutines in section 4.2.

We present our studies in section 4.9 for the theoretical parameters in table 4.2.

	Corresponding theoretical parameter		
(λ_L, λ_G)	π	z	R_*
(0.3, 0.12)	0.84487	0.3426	1.2902
(0.13, 0.17)	0.74223	0.4275	1.1432
(0.1, 0.29)	0.4199	0.7298	2.2166
(0.25, 0.39)	0.2302	0.9185	4.0229

Table 4.2: Pairs of the local and global infection rates with their corresponding theoretical parameters.

With household structure and population size fifty times that of [1] given as [133, 189, 108, 106, 31]$\times$ 50, minimum epidemic size of 1000, discussed in section 4.2 and simulation runs of 1000, in comparison with our studies in sections 4.5, 4.6 and 4.8 for theoretical parameters corresponding to $z = 0.1775$ and population size in [1] given in table 1.2, for different choice of the minimum epidemic size and simulation runs of 1000.

This is done in order to study the influence of the minimum epidemic size and the population size on the occurrence of a global infection in the households and hence the estimates of the parameters.

These are implemented using program function and subroutines in section 4.2. with the theoretical parameters in [1,9] and household structure in [1] with the population size of 1414, simulation runs of 1000 for the following minimum epidemic sizes $10, 50, 100$ respectively.

The scatter plots of the estimates and the histogram of the number infected are then presented to provide insights into their behaviours.

4.5 Plots of the estimates with minimum epidemic size of 10.

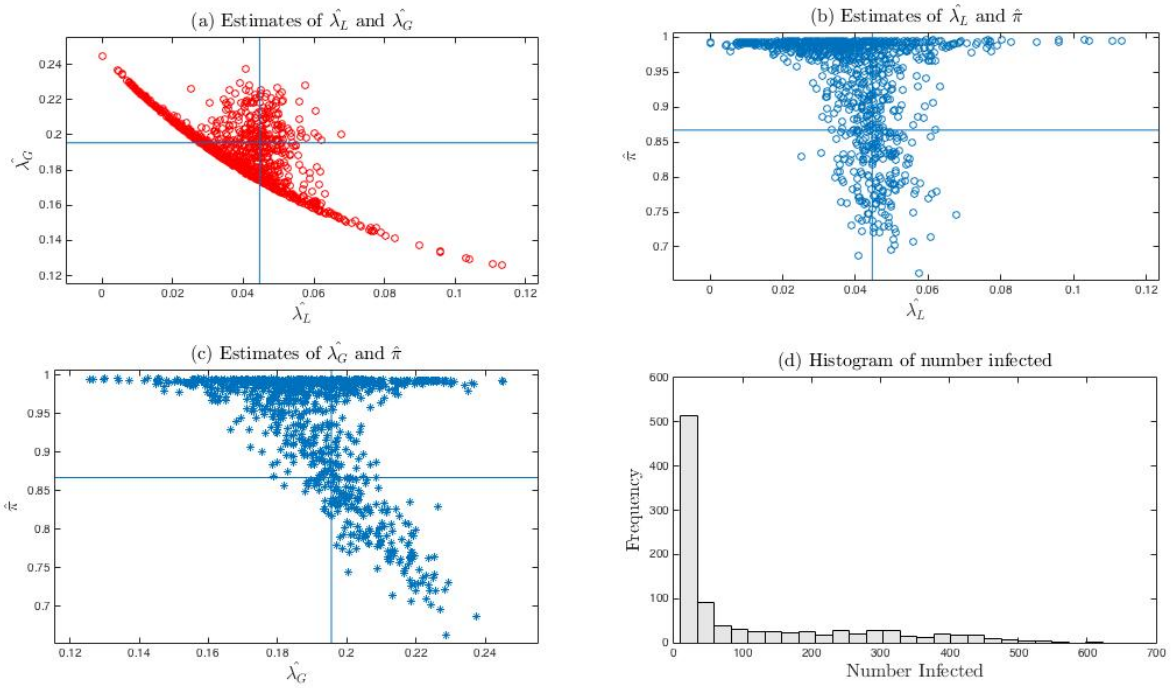

Figure 4.1: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and minimum epidemic size of 10.

In figure 4.1, we see positive and negative linear correlation between some of the parameter estimates for example increasing λ_L leads to decreasing λ_G. Generally, in most of the simulations few number of infections occurred, many suceptibles avoid global infection. Hence a global epidemic has not taken place.

4.5.1 Table of parameter estimates and other statistics when the minimum epidemic size is 10.

Mean, SD, MSE, RMSE.	Parameter Estimates.				
	$\hat{\lambda}_L$	$\hat{\lambda}_G$	$\hat{\pi}$	$\hat{z}$	$\hat{R}_*$
Theoretical Parameters	0.0446	0.1955	0.8674	0.1775	1.0596
Mean	0.038025	0.19346	0.94022	0.07902	1.0661
Standard Deviation	0.01582	0.020356	0.075164	0.098512	0.081201
Mean Square Error	0.00029238	0.00041811	0.010968	0.019431	0.011596
Root Mean Square Error	0.017125	0.020448	0.10473	0.1394	0.10769

Table 4.3: Mean of the parameter estimates for theoretical parameters corresponding to $z = 0.1775$, household structure and size in $[1, 9]$ and minimum epidemic size of 10.

In table 4.3, we see small difference between the mean of the estimates of λ_L, λ_G and their theoretical values. While those of π, z and R_* are significantly different from their theoretical mean and possess large standard deviation, which are the standard error of the estimates. These later three parameter estimates are biased owing to the choice of 10 as the minimum epidemic size with the small population size in [1].

The choice of the minimum epidemic size is further explored in section 4.6 and 4.8 to provide clarity on its effect on the parameters and the occurrence of a global infection in the households.

4.6 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 50.

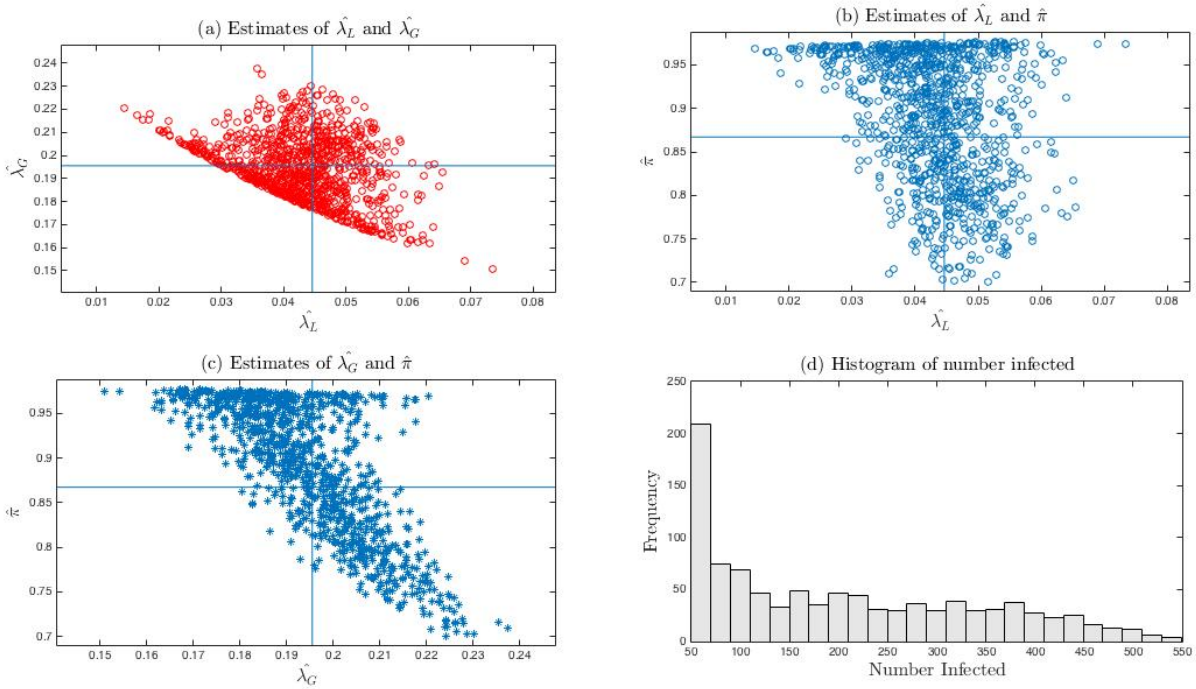

Figure 4.2: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and Minimum Epidemic size of 50.

In figures 4.2 (a)-(d), we see that the estimates are densely scattered around the true parameter values compared to the earlier case with minimum epidemic size of 10. Only few simulations resulted in large infections.

Most of the simulations yielded small number of infections, as many susceptibes avoided global infection.

4.6.1 Table of parameter estimates and other statistics when the minimum epidemic size is 50.

Mean, SD, MSE, RMSE.	Parameter Estimates.				
	$\hat{\lambda}_L$	$\hat{\lambda}_G$	$\hat{\pi}$	$\hat{z}$	$\hat{R}_*$
Theoretical Parameters	0.0446	0.1955	0.8674	0.1775	1.1303
Mean	0.043008	0.19459	0.88938	0.14652	1.1112
Standard Deviation	0.0083475	0.013859	0.073506	0.095575	0.082287
Mean Square Error	7.21E-05	0.0001927	0.0058872	0.010097	0.007133
Root Mean Square Error	0.008494	0.013882	0.076728	0.10048	0.084457

Table 4.4: Mean of the parameter estimates for theoretical parameters corresponding to $z = 0.1775$ and household structure and size in $[1, 9]$ and minimum epidemic size of 50.

In table 4.4, we see that the mean of the estimates of λ_L, and λ_G are approximately equal to their theoretical counterparts with the increase in the minimum epidemic size compared to those in table 4.3. The standard deviations and the mean square error are reasonably small.

The estimates are less biased compared to those in table 4.3. This indicates that appropriate choice of the minimum epidemic size leads to the realisation of a global infection in the households and hence the occurrence of a global epidemic in which there is enough information for the estimation of the parameters.

4.7 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 100.

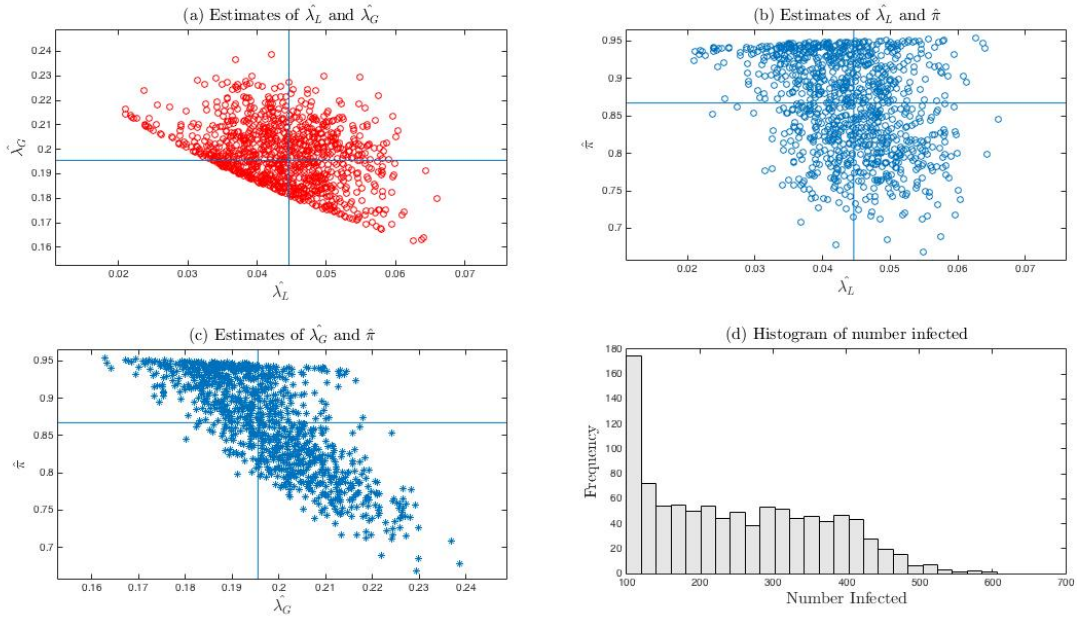

Figure 4.3: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters corresponding to $z = 0.1775$ and Minimum Epidemic size of 100.

In figures 4.3 (a)-(d), we see that the estimates are densely scattered around their true parameter values as in figures 4.2 (a)-(d) but with better precision and less bias as in table 4.5 compared to the earlier cases with minimum epidemic sizes of 10 and 50 respectively.

Also large number of simulations yielded few number infected with only small number of simulations with large number infected as shown by the bimodal behaviour of the histogram of the distribution of the number infected associated with simluations with small population size.

4.7.1 Table of parameter estimates and other statistics when the minimum epidemic size is 100.

Mean, SD, MSE, RMSE.	Parameter Estimates.				
	$\hat{\lambda}_L$	$\hat{\lambda}_G$	$\hat{\pi}$	$\hat{z}$	$\hat{R}_*$
Theoretical Parameters	0.0446	0.1955	0.8674	0.1775	1.1303
Mean	0.043791	0.19786	0.86498	0.17851	1.1362
Standard Deviation	0.0074468	0.012422	0.063977	0.082632	0.07399
Mean Square Error	5.61E-05	0.00015973	0.0040941	0.0068219	0.0055032
Root Mean Square Error	0.0074869	0.012638	0.063985	0.082595	0.074183

Table 4.5: Mean of the parameter estimates for theoretical paramters corresponding to $z = 0.1775$ and household structure and size in $[1, 9]$ and minimum epidemic size of 100.

In table 4.5, the estimates of the parameters are precise compared to those in tables 4.3 and 4.4. For example, λ_L, and λ_G are less biased with improved estimate.

4.8 Plots of the estimates and table of mean, standard deviation, mean square error and root mean square error with minimum epidemic size of 1000.

The behaviour of the estimates are further examined in table 4.6 as continuation of our studies with minimum epidemic sizes of 10, 50, and 100 in figures 4.1, 4.2, 4.3 with corresponding tables of statistics, 4.3, 4.4 and 4.5 respectively.

From table 4.5, we see that the estimates are unbiased given the population size in [1] and minimum epidemic size of 100 compared to the choice of minimum epidemic size less than 100. However, the question then is how precise are the estimates if the minimum epidemic size is extremely larger than 100, given the small population size of 1414 in [1] and also population size larger than 1414.

We explored these questions by assuming minimum epidemic sizes of 1000 for the small population size of 1414, which is far greater than 100, adopted in figures 4.5 (a)-(d). We employed the same minimum epidemic size of 1000 for the population of size of 70700, which

is fifty times greater than the population size considered in [1] as in table 4.6.

In the case of the small population size of 1414, a minimum epidemic size of 1000, give estimates that are biased and imprecise compared to the choice of 100 as the minimum epidemic size in table 4.5 with the same population size. Unlike in table 4.5, we see significant difference between the mean of the parameter estimates and their true values.

The mean square error of the estimates does not satisfy the minimum mean square error criterion required of good estimates. With large population size of, 70700, and choice of minimum epidemic size 1000, the estimates are unbiased with insignificant difference from their true mean values compared to the former as shown in table 4.6.

The choice of minimum epidemic size below and above its threshold given small and large population sizes affects the precision and other properties of the estimates of the parameters. Hence, there is the need to apply our discussion on the strategy of choosing this parameter in section 2.9. This involves, firstly simulating the household epidemic with minimum epidemic size of 1 to understand the bimodal behaviour of the distribution of the epidemic and hence locate the minimum cut-off of the number infected between the epidemics. Then use rejection sampling discussed in section 2.9.

From the bimodal behaviours of the distribution of the number infected in figure 4.4, for the small and large population sizes, 1414 and 70700, the cut-off of 100 and 1000 respectively are reasonable.

Choice of extremely large value above the minimum epidemic size leads to loss of information in the final size epidemic data. This is because simulations with large number infected will be rejected and hence may result in estimates that are biased and imprecise as shown in table 4.6, with minimum epidemic size of, 1000, for population sizes, 1414, and, 70700, respectively. The choice of, 1000, for the small population size of 1414, is far above the required cut-off between the epidemics as shown in figure 4.4 for small and large population sizes and hence some of the large epidemics will be wrongly rejected. This then leads to loss of information required for inference from the final size epidemic data. Hence biased estimates are obtained unlike the case with 100, in table 4.6.

Par.	Estim.	Pop. size=1414			Pop. size=70700		
		mean	std	MSE	mean	std	MSE
$\hat{\lambda_L}$	0.0446	0.053486	0.0089206	0.00015846	0.0445	0.0010809	1.18E-06
$\hat{\lambda_G}$	0.1955	0.33199	0.012481	0.018786	0.19525	0.0028492	8.17E-06
$\hat{\pi}$	0.86725	0.38183	0.013799	0.23583	0.86946	0.018014	0.00032903
$\hat{z}$	0.1777	0.70781	0.0013614	0.28103	0.17469	0.023642	0.00056745
$\hat{R_*}$	1.1304	2.0239	0.033412	0.7995	1.1282	0.019158	0.00037142

Table 4.6: Table of comparison of the mean, standard deviation and mean square error of the estimates using the minimum epidemic size of 1000 and simulation runs of 1000.

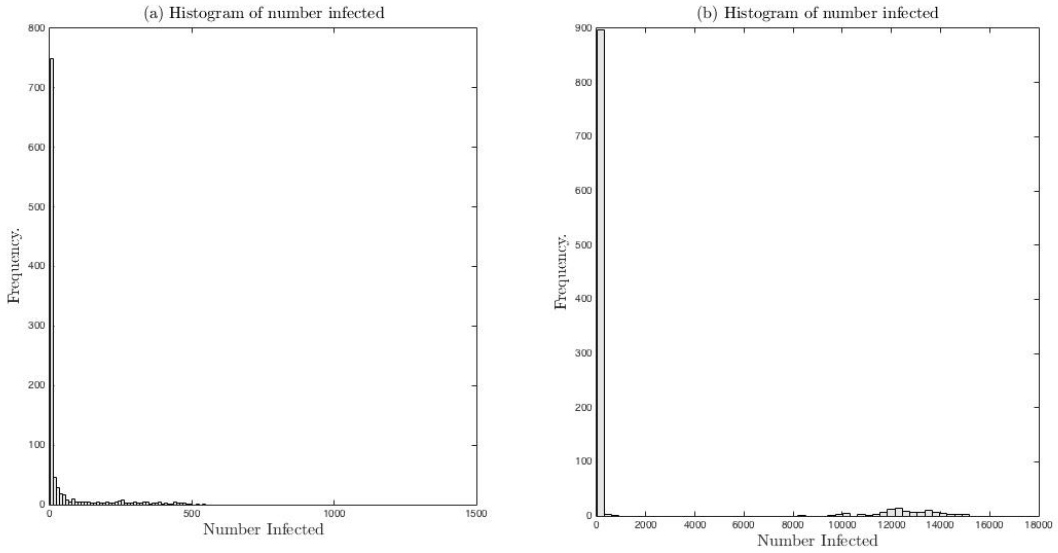

Figure 4.4: Histogram of number infected from simulations of household epidemic with population sizes of 1414 and 70700 respectively, minimum epidemic size of 1 and simulation runs of 1000.

4.9 Parameter estimates with minimum epidemic size of 1000.

In section 4.5 and 4.6, we considered in our simulation studies, small population size in [1] and minimum epidemic sizes of 10, 50 and 100 with the theoretical parameters in [1]. We found that in the face of varying minimum epidemic size, global infection failed to occur. Hence π, z and R_* are biased with imprecise estimates owing to lack of enough information in the final

size data. With increasing minimum epidemic size, these estimates become less biased with improved estimates.

In order to overcome this estimation problem, we considered large population size with appropriate minimum epidemic size of 1000 and a range of theoretical parameters in table 4.2 to allow global epidemic and hence provide sufficient information for parameter estimation.

We considered pair of theoretical parameters (λ_L, λ_G) corresponding to $0 < z < 0.5$ and $0.5 < z < 1$ away from its boundaries. We then studied the behaviour of the estimates and the distribution of the number infected for these sets of theoretical parameters corresponding to z in the given sets.

Starting with $\lambda_L = 0.0446$, $\lambda_G = 0.1955$ and corresponding theoretical parameters, $\pi = 0.8674$, $z = 0.1775$, $R_* = 1.1303$, minimum epidemic size of 1000, to allow global epidemic to take off in each of the simulation runs. We simulate 1000 times household epidemic in a population of size 70700 which is fifty times that of [1] given as 1414.

We then estimate and plot the parameters, (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of the distribution of number infected.

Table of mean, standard deviation and root mean square error of the estimates are presented.

4.9.1 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.0446$ and $\lambda_G = 0.1955$ with minimum epidemic size of 1000.

In figures 4.5 (a)-(d), the estimates are unbiased and scattered around their true parameter values. The unimodal pattern of the distribution of the number infected by the histogram is indicative of the occurrence of a global epidemics.

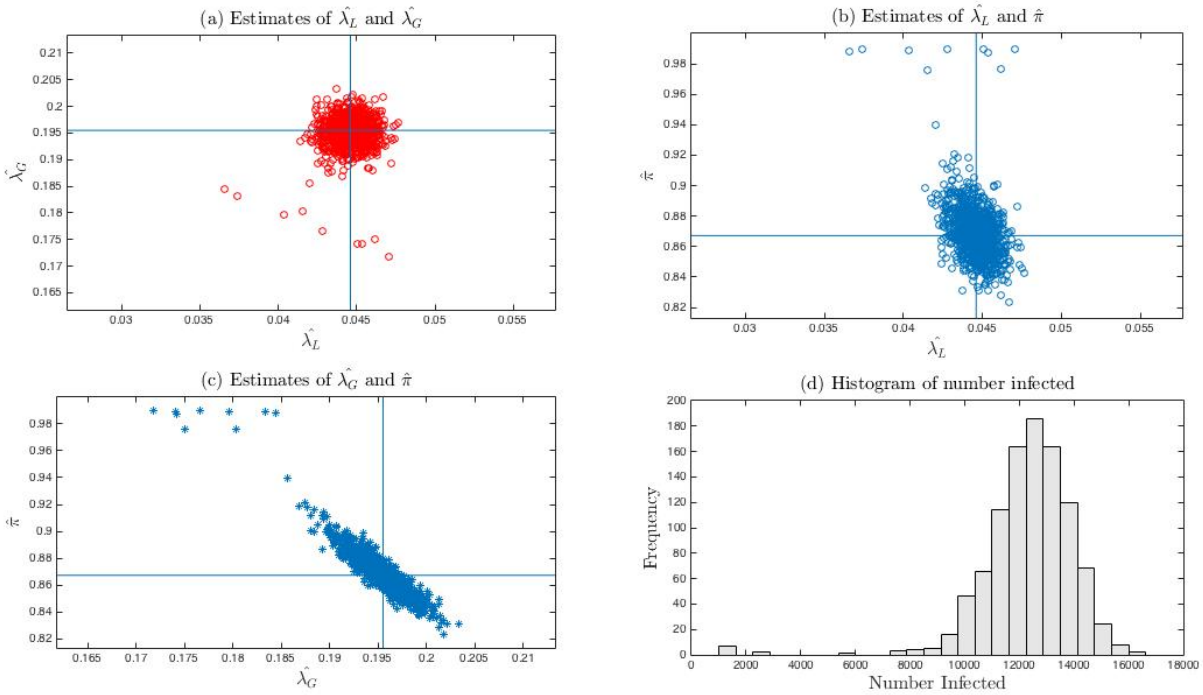

Figure 4.5: Plots of the Estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.0446$, $\lambda_G = 0.1955$ and minimum epidemic size of 1000.

4.9.2 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.13$ and $\lambda_G = 0.17$ with minimum epidemic size of 1000.

Now taking $\lambda_L = 0.13, \lambda_G = 0.17$, with corresponding theoretical parameters, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.1432$. We then estimate and plot (λ_L, λ_G), (λ_L, π), (λ_G, π) and the histogram of the distribution of number infected. Table of mean, standard deviation and root mean square error of the estimates are also presented.

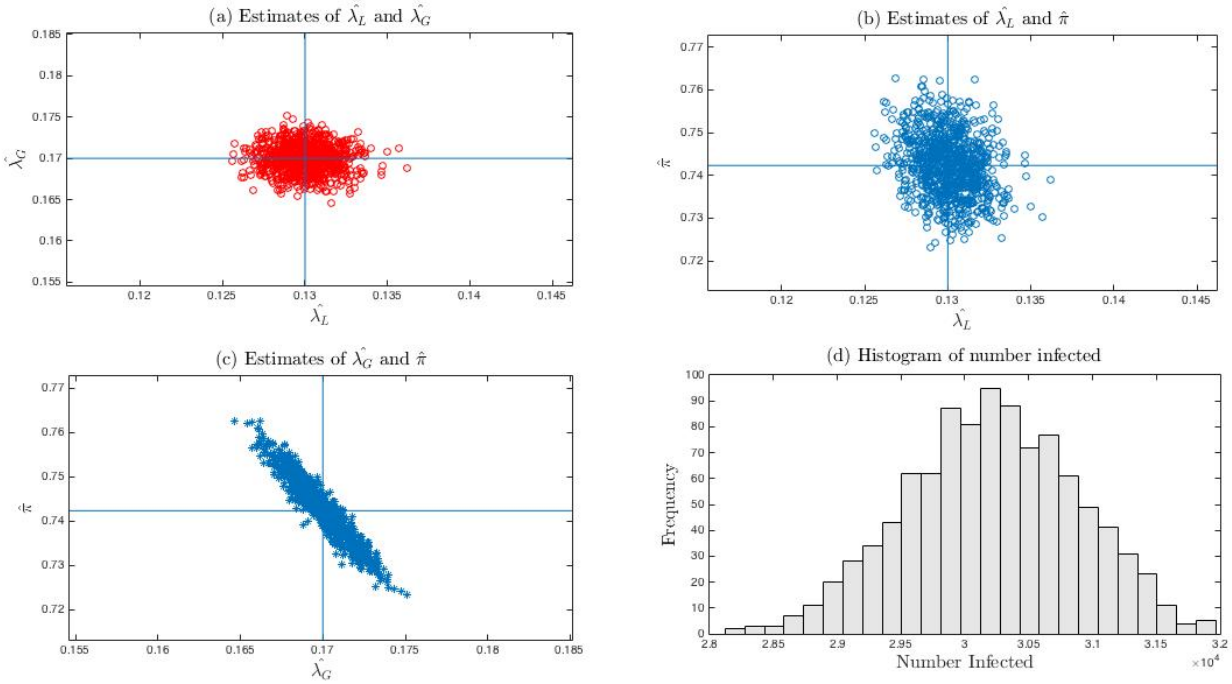

Figure 4.6: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.13, \lambda_G = 0.17$ and minimum epidemic size of 1000.

In figures 4.6 (a)-(d), similar behaviour to figures 4.5 (a)-(d) are seen, with linear correlation between λ_G and π and good precision and more number of susceptibles are infected.

4.9.3 Plots of the estimates of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.1$ and $\lambda_G = 0.29$ with minimum epidemic size of 1000.

Similarly, we simulated household epidemic with $\lambda_L = 0.1, \lambda_G = 0.29$ and corresponding theoretical parameters, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$ and plotted (λ_L, λ_G), (λ_L, π), (λ_G, π) and the histogram of the distribution of number infected in figures 4.7 (a)-(d). Table of mean, standard deviation and root mean square error of the estimates also provided.

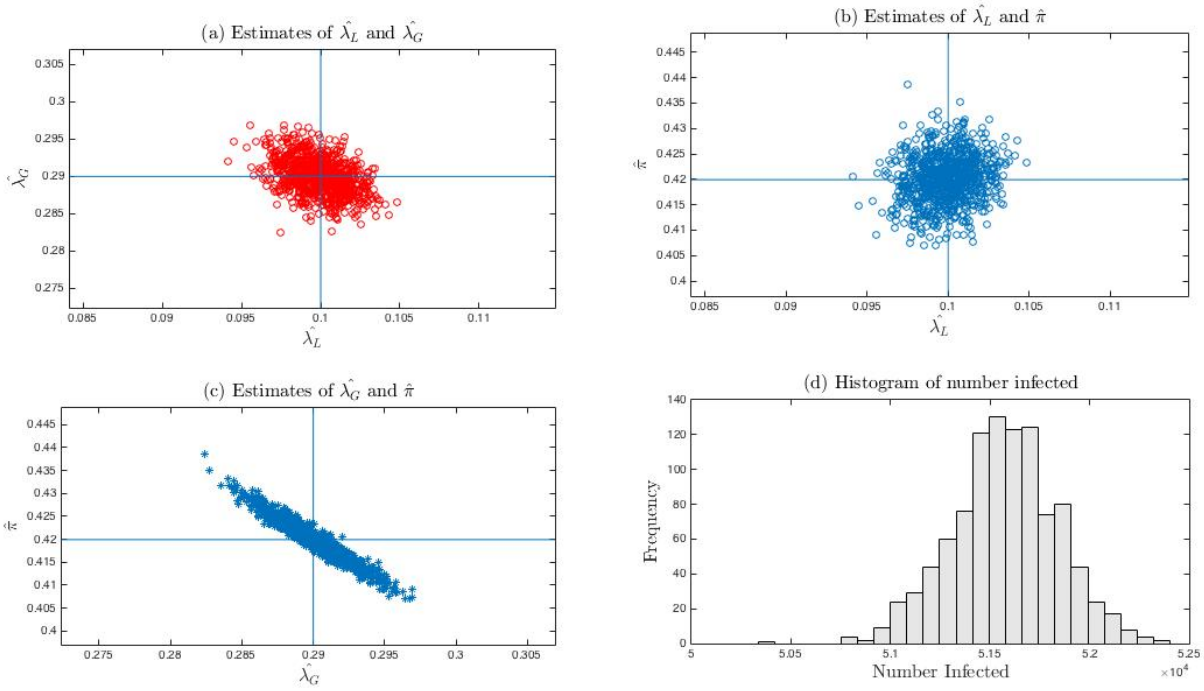

Figure 4.7: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.1, \lambda_G = 0.29$ and minimum epidemic size of 1000.

In figures 4.7 (a)-(b), similar behaviours of the estimates in figures 4.6 and 4.5 are shown with the scatter points around their true value. Large number of susceptibles are infected.

4.9.4 Plots of the estimate of λ_L, λ_G and π when the theoretical parameters are $\lambda_L = 0.25$ and $\lambda_G = 0.39$ with minimum epidemic size of 1000.

Also with $\lambda_L = 0.25, \lambda_G = 0.39$ and corresponding theoretical parameters, $\pi = 0.2302$, $z = 0.9185$, $R_* = 4.0229$. We plotted (λ_L, λ_G), (λ_L, π), (λ_G, π) and the histogram of the distribution of number infected. Table of mean, standard deviation and root mean square error are presented.

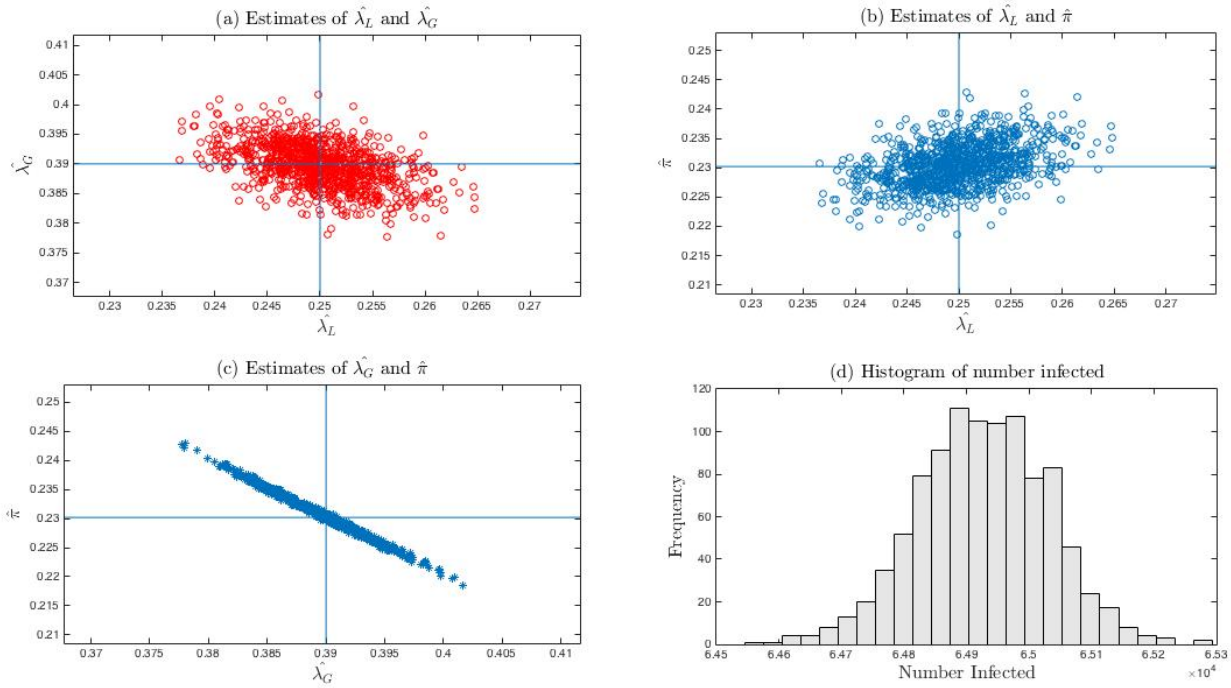

Figure 4.8: Plots of the estimates of (λ_L, λ_G), (λ_L, π), (λ_G, π) and histogram of number infected with theoretical parameters $\lambda_L = 0.25, \lambda_G = 0.39$ and minimum epidemic size of 1000.

In figures 4.8 (a)-(d), the estimates are precise and centred around the true parameter values. Also large number of susceptibles are infected.

Par.	Proportion Infected.							
	z=0.1775	Theor. Par.	z=0.42757	Theor. Par.	z=0.7298	Theor. Par.	z=0.9185	Theor. Par.
$\hat{\lambda_L}$	0.044578	0.0446	0.13004	0.13	0.099901	0.1	0.24987	0.25
$\hat{\lambda_G}$	0.19515	0.1955	0.16997	0.17	0.28997	0.29	0.38983	0.39
$\hat{\pi}$	0.86956	0.8674	0.74247	0.7423	0.42011	0.4199	0.23046	0.23021
$\hat{z}$	0.17461	0.1775	0.42728	0.42757	0.72949	0.7298	0.91833	0.9185
$\hat{R_*}$	1.1282	1.1303	1.4315	1.4316	2.2154	2.2166	4.0203	4.0229

Table 4.7: Table of mean of the estimates from the two dimensional model and theoretical parameters in table 4.2.

Par.	Proportion Infected.			
	z=0.1775	z=0.42757	z=0.7298	z=0.9185
$\hat{\lambda_L}$	0.0010624	0.0015197	0.0015715	0.0047053
$\hat{\lambda_G}$	0.0030219	0.0016325	0.0023247	0.0036795
$\hat{\pi}$	0.018571	0.006892	0.0045573	0.0036487
$\hat{z}$	0.024377	0.0094885	0.0037947	0.0014917
R_*	0.019749	0.014713	0.017152	0.033281

Table 4.8: Table of the standard deviation of the estimates from the two dimensional model with theoretical parameters in table 4.2

Parameter.	Proportion Infected.			
	z=0.1775	z=0.42757	z=0.7298	z=0.9185
$\hat{\lambda_L}$	0.0010621	0.0015196	0.0015738	0.0047048
$\hat{\lambda_G}$	0.0030408	0.001632	0.0023238	0.0036814
$\hat{\pi}$	0.018705	0.0068906	0.004558	0.0036554
$\hat{z}$	0.024559	0.009487	0.0037996	0.0015001
$\hat{R_*}$	0.019861	0.014707	0.017183	0.033367

Table 4.9: Table of the root mean square error of the estimates from the two dimensional model with theoretical parameters in table 4.2. The estimates are precise.

Chapter 5

Stochastic SIR household model for misclassified data.

5.1 Introduction

Mismeasurement of individual health state can be expressed in terms of misclassification probabilities, defined as the probability of classifying a subject into group i while its true status is in j. This leads to imprecise records of the number of individuals infected in each household and therefore unreliable results of inferences from such data. It then becomes necessary to adjust our inferences to such errors to get the appropriate parameter estimates and model that represents our data.

In this chapter, we present the theoretical basis leading to identification and estimation of classification error probability of the SIR household epidemic model. Its influence on the maximum likelihood estimates of the parameters is studied using simulations.

Mismeasurement occur when infectives are wrongly classified as susceptibles or susceptibles classified as infectives. Here we have assumed these misclassification probabilities to be independent and different from each and also examined the particular case when they are the same in section 5.3.

In section 5.2, we developed the theoretical basis leading to the four dimensional model with different misclassification probabilities and then extended it to the case with the same

misclassification probabilities. We discussed its estimation procedures using [1] maximum likelihood algorithm. Parameter estimations are implemented in section 5.5, using codes developed during this research. It computes the mean, standard deviation, root mean square error of the estimates and plots the estimates.

In section 5.6, we explored the parameter estimates of the four dimensional model along the vertical, horizontal axes and along the diagonals of the misclassification probabilities region, $\varepsilon \in [0, 0.2)$, with step size of 0.005 and theoretical parameters, $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.1432$ and those for $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, with $R_* = 2.2166$ for the three models using simulation studies.

Plots of the root mean square error of the estimates of the three models are presented in order to provide insight into their precision over the misclassification probability region.

We also presented table of comparison of the model estimates for misclassification probabilities in $[0, 0.2]$. In section 5.7, we discussed the behaviours of the three models on data from the four dimensional model for misclassification probabilities in the permissible region, $[0, 0.5]$.

In section 5.4 we explored the estimates of the two and three dimensional models for a range of misclassification probabilities in the permissible region, $[0, 0.5]$. We compute the mean, standard deviation, root mean square error for the two models and also plot their root mean square error for $\varepsilon \in [0, 0.5]$.

5.2 The SIR household epidemic model with two different misclassification probabilities.

We have assumed that the stochastic SIR household final size data is subject to misclassification error; which may be caused by susceptibles wrongly classified as infectives or infectives wrongly classified as susceptibles. The probability of making these classification errors are referred to as false negative and positive probabilities denoted here by ε_{FN} and ε_{FP} respectively.

The probability of observing i infectives in a household of size n given that the true number

of infectives is j and that of the susceptibles is $n-j$ takes cognisance of the true and false positives with their classification probabilities $1-\varepsilon_{FN}$ and ε_{FP}.

Let x and y be the observed false and true positives in a household of size n. Then the probability of observing $x+y=i$ positives, given that the true number of positives is j can be written as,

$$P_{i,j}(n) = P(x+y=i \,|\, \text{True infect} = j, \text{ household size} = n). \tag{5.2.1}$$

Using the false positive and false negative probabilities, we can express, the probability of making correct and precise observation of an infective when it is a true infective, and a susceptible, when it is a true susceptible, independently as, $1-\varepsilon_{FN}$ and $1-\varepsilon_{FP}$. The distribution of observing i number of infectives correctly and incorrectly is Binomial distributed, $\text{Bin}(j, 1-\varepsilon_{FN})$, and $\text{Bin}(n-j, \varepsilon_{FP})$. Equally the probability of observing the susceptibles correctly and incorrectly are Binomial distributed, $\text{Bin}(n-j, 1-\varepsilon_{FP})$ and $\text{Bin}(j, \varepsilon_{FN})$ respectively.

The number of infectives observed is the sum of the true and false positives and has the sum of the Binomial distributions,

$$Bin(j, 1-\varepsilon_{FN}) + Bin(n-j, \varepsilon_{FP}). \tag{5.2.2}$$

Equally, the number of susceptibles observed is the sum of the true and false negatives and has the sum of the Binomial distributions,

$$Bin(n-j, 1-\varepsilon_{FP}) + Bin(j, \varepsilon_{FN}). \tag{5.2.3}$$

The probability of observing i infectives in a household of size of n can then be written as,

$$q_{n,i} = \sum_{j=0}^{n} P(\text{Obs} = i, \text{ True infect} = j, \text{ household size} = n). \tag{5.2.4}$$

Since,

$$P(\text{Obs} = i,\ \text{True} = j,\ \text{household size} = n)$$
$$= P(\text{Obs} = i \,|\, \text{True} = j,\ \text{household size} = n)P(\text{True} = j). \tag{5.2.5}$$

$$q_{n,i} = \sum_{j=0}^{n} P(x + y = i \,|\, \text{True infect} = j,\ \text{household size} = n)P(\text{True} = j), \tag{5.2.6}$$

where P(True=j), $j = 0, 1, \ldots, n$ are the final size probabilities described in equations (2.6.2) and (2.6.3). We can then write,

$$q_{n,i} = \sum_{j=0}^{n} P_{i,j}(n)P_j(n),\ i = 0, 1, \cdots, n. \tag{5.2.7}$$

Where

$$P_{i,j}(n) = P(x + y = i \,|\text{True} = j,\ \text{household size} = n). \tag{5.2.8}$$

For example, we can evaluate the terms of $P_{i,j}(n)$ starting with $i = 0$ as,

$$P_{0,0}(n) = P(x + y = 0 \,|\, j = 0, \text{household size} = n)$$
$$= P(x = 0 \,|\, j = 0)P(y = 0 \,|\, j = 0) = (1 - \varepsilon_{FP})^n$$
$$P_{0,1}(n) = P(x + y = 0 \,|\, j = 1) = P(x = 0 \,|\, j = 0)P(y = 0 \,|\, j = 1)$$
$$= \varepsilon_{FN}(1 - \varepsilon_{FP})^{n-1}$$
$$P_{0,2}(n) = p(x + y = 0 \,|\, j = 2, \text{ household size} = n)$$
$$= P(x = 0 \,|\, j = 2, \text{ household size} = n)P(y = 0 \,|\, j = 2, \text{ household size} = n)$$
$$= \varepsilon_{FN}^2(1 - \varepsilon_{FP})^{n-2}$$
$$P_{0,3}(n) = \varepsilon_{FN}^3(1 - \varepsilon_{FP})^{n-3}$$
$$\cdots = \cdots$$
$$P_{0,j}(n) = \varepsilon_{FN}^j(1 - \varepsilon_{FP})^{n-j},\ j = 0, 1, \cdots, n. \tag{5.2.9}$$

Also the probability of observing $i = 1$ infective in a household of size n, given that the true

number of infectives is $j = 0, 1, \cdots, n$ can be evaluated as follows,

$$P_{1,0}(n) = P(x + y = 1 \,|\, j = 0, \text{ household size} = n)$$

$$= P(x = 1 \,|\, j = 0)P(y = 0 \,|\, j = 0) + P(x = 0 \,|\, j = 0)P(y = 1 \,|\, j = 0) = n\varepsilon_{FP}(1 - \varepsilon_{FP})^{n-1}$$

$$P_{1,1}(n) = P(x + y = 1 \,|\, j = 1, \text{ household size} = n)$$

$$= P(x = 1 \,|\, j = 1, \text{ household size} = n)P(y = 0 \,|\, j = 1, \text{ household size} = n)$$

$$+P(x = 0 \,|\, j = 1, \text{ household size} = n)P(y = 1 \,|\, j = 1, \text{ household size} = n)$$

$$= (n - 1)\varepsilon_{FP}\varepsilon_{FN}(1 - \varepsilon_{FP})^{n-2} + (1 - \varepsilon_{FN})(1 - \varepsilon_{FP})^{n-1}$$

$$P_{1,2}(n) = P(x + y = 1 \,|\, j = 2, \text{ household size} = n) =$$

$$P(x = 0 \,|\, j = 2, \text{ household size} = n)P(y = 1 \,|\, j = 2, \text{ household size} = n)$$

$$+P(x = 1 \,|\, j = 2, \text{ household size} = n)P(y = 0 \,|\, j = 2, \text{ household size} = n)$$

$$= (n - 2)\varepsilon_{FN}^2\varepsilon_{FP}(1 - \varepsilon_{FP})^{n-3} + 2\varepsilon_{FN}(1 - \varepsilon_{FN})(1 - \varepsilon_{FP})^{n-2}$$

$$\cdots = \cdots$$

$$P_{1,j}(n) = P(x + y = 1 \,|\, \text{Truth} = j, \text{ household size} = n) =$$

$$(n - j)\varepsilon_{FN}^j\varepsilon_{FP}(1 - \varepsilon_{FP})^{n-j-1} + j\varepsilon_{FN}^{j-1}(1 - \varepsilon_{FN})(1 - \varepsilon_{FP})^{n-j},\ j = 0, 1, \cdots, n.$$

Thus,

$$P_{2,j}(n) = j(n - j)\varepsilon_{FP}(1 - \varepsilon_{FP})^{n-j-1}(1 - \varepsilon_{FN})\varepsilon_{FN}^{j-1} + \frac{(n - 1)(n - j - 1)}{2!}\varepsilon_{FN}^j$$

$$+\varepsilon_{FP}^2(1 - \varepsilon_{FP})^{n-j-2} + \frac{j(j - 1)}{2!}(1 - \varepsilon_{FN})^2\varepsilon_{FN}^{j-2}(1 - \varepsilon_{FP})^{n-j},\ j = 0, 1, \cdots, n$$

$$P_{3,j}(n) = \frac{(n-j)(n-j-1)(n-j-2)}{3!}\varepsilon_{FP}^{3}\varepsilon_{FN}^{j}(1-\varepsilon_{FP})^{n-j-3}$$
$$+\frac{j(j-1)(j-2)}{3!}(1-\varepsilon_{FN})^{3}\varepsilon_{FN}^{j-3}(1-\varepsilon_{FP})^{n-j}$$
$$+\frac{j(n-j)(n-j-1)}{2!}(1-\varepsilon_{FN})\varepsilon_{FN}^{j-1}\varepsilon_{FP}^{2}(1-\varepsilon_{FP})^{n-j-2}$$
$$+\frac{j(j-1)(n-j)}{2!}(1-\varepsilon_{FN})^{2}\varepsilon_{FN}^{j-2}\varepsilon_{FP}(1-\varepsilon_{FP})^{n-j-1}$$

When $j = n$ which is the household size then,

$$P_{3,n}(n) = \frac{n(n-1)(n-2)}{3!}(1-\varepsilon_{FN})^{3}\varepsilon_{FN}^{n-3},\ n \geq 3$$

It is more useful to generalise the expression, $P_{i,j}(n)$ for $i,\ j = 0, 1, 2\cdots, n$ and any $r \in \mathbb{Z}_{+} \leq n$ using the results of $P_{0,j}(n), P_{1,j}(n)\cdots, P_{i,j}(n)$ as follows,

$$P_{i,j}(n) = P(x+y=i \,|\, \text{Truth} = j,\ \text{household size} = n)$$
$$= P(x=0 \,|\text{Truth} = j,\ \text{household size} = n)P(y=i \,|\, \text{Truth} = j,\ \text{household size} = n)$$
$$+P(x=1 \,|\, \text{Truth} = j,\ \text{household size} = n)P(y=i-1 \,|\, \text{Truth} = j,\ \text{household size} = n)$$
$$+P(x=2 \,|\, \text{Truth} = j,\ \text{household size} = n)P(y=i-2 \,|\, \text{Truth} = j,\ \text{household size} = n)$$
$$+\cdots+P(x=i-1 \,|\, \text{Truth} = j,\ \text{household size} = n)P(y=1 \,|\, \text{Truth} = j,\ \text{household size} = n)$$
$$P(x=i \,|\, \text{Truth} = j,\ \text{household size} = n)P(y=0 \,|\, \text{Truth} = j,\ \text{household size} = n)$$

$$P_{i,j}(n) = \frac{j(j-1)(j-2)\cdots(j-i+1)}{r!}\varepsilon_{FN}^{j-i}(1-\varepsilon_{FN})^{r}(1-\varepsilon_{FP})^{n-j}$$
$$+\frac{j(j-1)\cdots(j-i+2)}{(i-1)!}(n-j)\varepsilon_{FP}(1-\varepsilon_{FP})^{n-j-1}\varepsilon_{FN}^{j-i+1}(1-\varepsilon_{FN})^{i-1}$$
$$+\frac{j(j-1)(j-2)\cdots(j-i+3)}{(i-2)!}\frac{(n-j)(n-j-1)}{2!}\varepsilon_{FP}^{2}(1-\varepsilon_{FP})^{n-j-2}\varepsilon_{FN}^{j-i+2}(1-\varepsilon_{FN})^{i-2}$$
$$+\frac{(j(j-1)(j-2)\cdots(j-i+4)}{(r-3)!}\frac{(n-j-1)(n-j-2)}{3!}\varepsilon_{FP}^{3}(1-\varepsilon_{FP})^{n-j-3}(1-\varepsilon_{FN})^{i-3}\varepsilon_{FN}^{j-i+3}$$
$$+\cdots+$$

$$\frac{(n-j)(n-j-1)\cdots(n-j-i+2)}{(r-1)!}\varepsilon_{FP}^{i-1}(1-\varepsilon_{FP})^{n-j-i+1}j(1-\varepsilon_{FN})\varepsilon_{FN}^{j-1}$$
$$+\frac{(n-j)(n-j-1)\cdots(n-j-i+1)}{r!}\varepsilon_{FP}^{i}(1-\varepsilon_{FP})^{n-j-i}\varepsilon_{FN}^{j} \tag{5.2.10}$$

Knowing the terms of $P_{i,j}(n), i, j = 0, 1, \cdots, n$, the expression for $q_{n,i}, i = 0, 1, \cdots, n$ can be evaluated. For example the probability of observing $i = 0$ infectives in a household of size n can be evaluated using equation(5.2.7) as,

$$q_{n,0} = \sum_{j=0}^{n} P_{0,j}(n)P_j(n),\ j = 0, 1, \cdots, n.$$

Where $P_j(n)$ are the final size probabilities, defined as the probabilitity of observing j infectives in a household of size n. We can then evaluate $P_{0,j}(n)$ from equation (5.2.9) for all $j = 0, 1, \cdots, n$.

Similarly, the chance of observing $i = 1$ infectives in a household of size n can be obtained using the terms of $P_{1,j}(n)\,\forall j \in \mathbb{Z}_+ \leq n$. This probability reduces to,

$$q_{n,1} = \sum_{j=0}^{n} P_{1,j}(n)P_j(n)$$

In general, the probability of observing $i \in \mathbb{Z}_+ \leq n$ infectives in a household of size n, is similarly obtained as,

$$P_{r,j}(n) = \sum_{k=0}^{r} \binom{j}{r-k}\binom{n-j}{k}\varepsilon_{FN}^{j-r+k}(1-\varepsilon_{FN})^{r-k}\varepsilon_{FP}^{k}(1-\varepsilon_{FP})^{n-j-k} \tag{5.2.11}$$

Equations (5.2.11) is the sum of two Binomial distributions, $\text{Bin}(j, (1-\varepsilon_{FN}))$ and $\text{Bin}(n-j, \varepsilon_{FP})$ defined as the probabilities of observing $r-k$ true positives from the true j number of infectives and k false positives from the remaining $n-j$ number of susceptibles in a household of size n.

Alternatively, $P_{r,j}(n)$ has the form,

$$P_{r,j}(n) = \sum_{k=0}^{r} \binom{j}{k} \binom{n-j}{r-k} \varepsilon_{FN}^{j-k} (1-\varepsilon_{FN})^{k} \varepsilon_{FP}^{r-k} (1-\varepsilon_{FP})^{n-j-r+k}. \tag{5.2.12}$$

Equation (5.2.12) is also the sum of two Binomial distributions in equation (5.2.11) and defined as the probability of observing k true positives from the true j infectives and $r-k$ false positives from the remaining $n-j$ susceptibles in a household of size n.

Here, both equations (5.2.11) and (5.2.12) for $P_{r,j}(n)$ satisfy,

$$\sum_{i=0}^{n} P_{i,j}(n) = 1, \, \forall j \in \mathbb{Z}_{+} \leq n.$$

5.3 The three dimensional final size epidemic model.

If the false positive and false negative misclassification probabilities are the same then equations (5.2.2) and (5.2.3) for the distribution of the number of infected individuals observed and those of the susceptible individuals observed only depend on the common misclassification probability denoted here as ε. In these equations, ε_{FN} and ε_{FP} are replaced by ε same as in the expressions for $P_{i,j}(n)$, $i, j = 0, 1, \ldots, n$ and simplified as,

$$P_{i,j}(n) = \sum_{k=0}^{i} \binom{j}{i-k} \binom{n-j}{k} \varepsilon^{j-i+2k} (1-\varepsilon)^{n-j+i-2k}, \quad i, j = 0, 1, \ldots, n. \tag{5.3.1}$$

Alternatively, we can employ

$$P_{i,j}(n) = \sum_{k=0}^{i} \binom{j}{k} \binom{n-j}{i-k} \varepsilon^{j+i-2k} (1-\varepsilon)^{n-j-i+2k}, \quad i, j = 0, 1, \ldots, n. \tag{5.3.2}$$

Equations (5.3.1) and (5.3.2) for $P_{i,j}(n)$ which are particular cases of equations (5.2.11) and (5.2.12) when the misclassification probabilities are the same are made of two Binomial distributions. While equation (5.3.1) expresses the probability of observing $i-k$ infectives from

the true j infectives and k infectives from the remaining $n-j$ susceptibles in the household of size n, equation (5.3.2) expresses the probability of observing k infectives from the true j infectives and $i-k$ infectives from the remaining $n-j$ susceptibles in the household of size n.

Since they are probabilities, both equations $P_{i,j}(n)$, must satisfy,

$$\sum_{i=0}^{n} P_{i,j}(n) = 1,\ \forall j \in \{0, 1, \ldots, n\}.$$

5.3.1 Maximum likelihood estimation.

In section 2.12, we see that the distribution of the final size epidemic data $x_{n,i}$ is multinomial, where $x_{n,i}$ are the number of households of size n in which i infectives are observed and $q_{n,i}$ are the probabilities of observing i infectives in a household of size n. The approximate likelihood function of the model parameters is then a function of $q_{n,i}$ and dependent on the parameters to be estimated from the four dimensional model. These parameters are the local infection rate λ_L, the probability of avoiding infection from outside the household π, the false positive misclassification probability, ε_{FP} and the false negative misclassification probability, ε_{FN} and hence $q_{n,i}$ has the form $q_{n,i}(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN})$. The approximate likelihood function discussed in section 4.2 then has the form,

$$L(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN}) \propto \prod_{n=1}^{max} \prod_{i=0}^{n} q_{n,i}(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN})^{x_{n,i}}. \tag{5.3.3}$$

where max is the maximum household size.

Since the estimates that maximize the approximate likelihood function also maximize the approximate loglikelihood function, we can write,

$$\ell(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN}) = \sum_{n=1}^{max} \sum_{i=0}^{n} \left(x_{n,i} \log_e \left(\sum_{j=0}^{n} P_{i,j}(n) P_j(n) \right) \right),\ i,\ j = 0, 1, \cdots, n. \tag{5.3.4}$$

Where $\log(L(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN})) = \ell(\lambda_L, \pi, \varepsilon_{FP}, \varepsilon_{FN})$

The approximate likelihood function for the three dimensional model also has similar rep-

resentation to that of the four dimensional model with differences in the number of parameters to be estimated.

5.4 Numerical simulations and inferences on the three and four dimensional final size epidemic data.

How precise are the maximum likelihood estimates from the numerical optimizations, given the minimum epidemic and population sizes, the proportion of the initial susceptibles infected and the magnitude of the misclassification probabilities? Which of these parameters are intractable to estimate in the face of large misclassification probabilities? Which model best fits the final size epidemic data in the face of varying misclassification probabilities in the permissible region, $[0, 0.5)$? These are some of the questions to be explored in this section using simulation studies. The term minimum epidemic size has been discussed in sections 4.2, while two, three and four dimensional final size epidemic data can also be found in section 1.10.

5.4.1 Fitting the three models to data from the four dimensional model.

Here, we have demonstrated the computational procedures of fitting the three models to four dimensional epidemic data from simulation studies and then studied the behaviours of the estimates using the following function and subroutines.

Run the function FourDimThreeATwoSNsimhousesScatterPlotsMisspec to simulate four dimensional household epidmemic data with Gamma(a, b) infectious period distribution, theoretical parametes, λ_L, λ_G and ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$. It then calculate the corresponding parameters of the three models with Gamma(a, b) infectious period distribution computes, their mean, standard deviation and root mean square error of the estimates and plot the estimates using the following subroutines.

a.) LampaiD(mat), provides starting values for the two dimensional model parameters, λ_L and π according to [24].

b.) Enegloglik4(y, n, a, b, mat), computes the negative of the loglikelihood function as-

sociated with the three dimensional model using the parameters of Gamma(a, b) infectious period distribution, the final size epidemic data and the starting parameters values obtained by inverse transformation of the parameter space.

c.) negloglik2(x, n, a, b, mat), computes the negative loglikelihood function associated with the two dimensional model from the parameters of Gamma(a, b) infectious period distribution, the final size epidemic data and the starting values according to [24].

d.) Misclass2(ε, n), computes the misclassification Probabilities asssociated with the three dimensional model from the misclassification probability parameter ε and maximum household size n

e.) final_sizep$(a, b, \pi, n, \lambda_L)$ computes the final size probabilities associated with the two dimensional model from the parameters of Gamma(a, b) infectious period distribution, π, λ_L and maximum household size n.

f.) Misclass3$(a, b, n, \pi, \lambda_L, \varepsilon)$, computes the sum of the product of the misclassification probabilities and the final size probabilities associated with the three dimensional model for the computation of the negative loglikelihood function.

g.) falseMisclass2$(\varepsilon_{FN}, \varepsilon_{FP}, n)$, computes the misclassification probabilities asssociated with the four dimensional model.

h.) SIRfalsePmisclass$(a, b, n, \pi, \lambda_L, \text{fneg}, \text{fpos})$, computes the products of the misclassification probabilities and the final size probabilities associated with the loglikelihood function of the four dimensional model.

i.) pinf2$(a, b, \pi, \lambda_L \text{houses})$, calculates z and λ_G, from the parameters of Gamma(a, b) infectious period distribution, model parameters π, λ_L and vector of household sizes, where houses is the vector of household sizes.

j.) RSTER2$(a, b, c, \lambda_L, \lambda_G, houses)$ calculates the threshold parameter, R_* from the parameters of Gamma(a, b) infectious period distribution, theoretical parameters λ_L, λ_G and vector of household sizes, houses.

Using the theoretical parameters, $z = 0.7298$, $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$, household structure in [1] but fifty times its population size given by 70700, minimum epidemic size of 1000 and simulation runs of 1000. The estimates of the parameters of the

three models were obtained for the following pairs of the misclassification probabilities ($\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$), ($\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$) and ($\varepsilon_{FN} = 0.2, \varepsilon = 0.2$) respectively shown in figures 5.1, 5.2 and 5.3 and analysed in tables 5.1, 5.2 and 5.3 respectively.

We observed that with large misclassification probabilities, $\varepsilon_{FN}, \varepsilon_{FP}$ in the permissible region, the estimates of the two dimensional model are imprecise and biased, while those of the three dimensional models are less precise. Better precision of the estimates can be seen from those of the four dimensional model in table 5.1 and figures 5.1 (a), (b) and (c). In figures 5.2 (a), (b) and (c), the estimates from the two and three dimensional models are biased and imprecise while those from the four dimensional model are unbiased and precise.

With the false negative and false positive misclassification probabilities assumed to be the same in figures 5.3 (a), (b) and (c), we see that both the three and four dimensional models have unbiased and precise estimates compared to those from the two dimensional model in figure 5.3 (c).

In general, the estimates of the four dimensional model have higher level of precision than those of the three dimensional models when the misclassification probabilities are large and far apart while those of the two dimensional are biased and imprecise.

Thus the four dimensional model outperforms the two and three dimensional models on the four dimensional final size epidemic data.

These model estimates are further explored in section 5.6 for varying values of the misclassification probabilities in the region, $[0, 0.2)$.

5.4.2 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$.

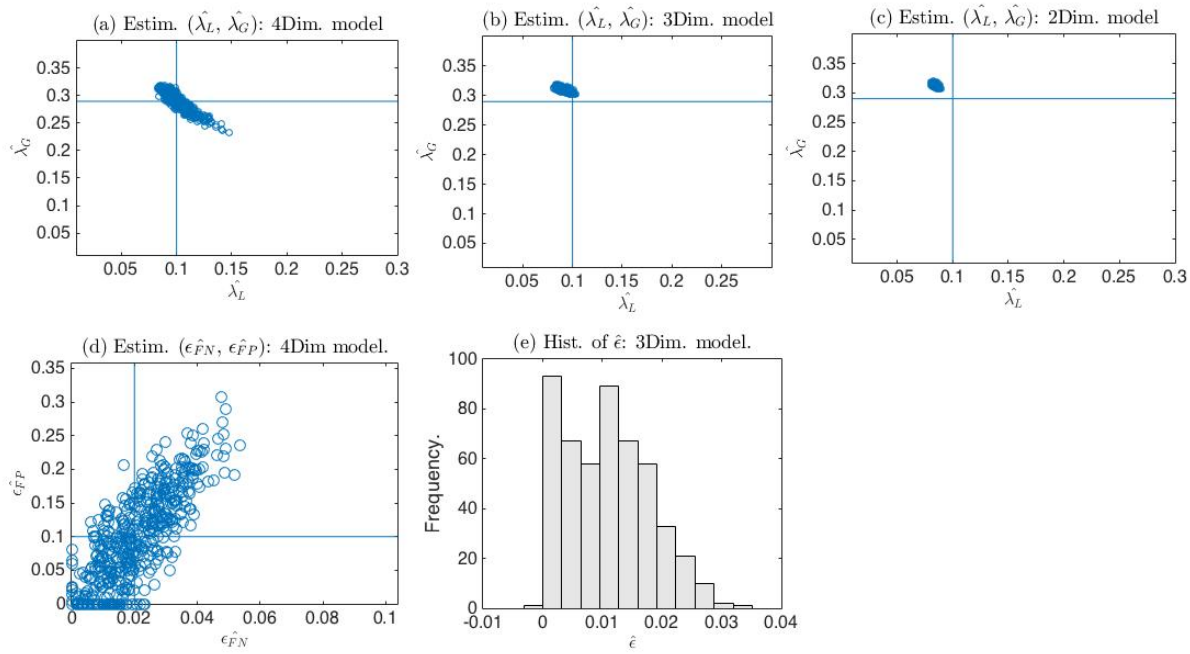

Figure 5.1: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$.

In figures 5.1 (a), (b) and (c), we see that the estimates of the local and global infection rates from the two and three dimensional models are biased, while those of the four dimensional models have more variability around the true values.

5.4.3 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$.

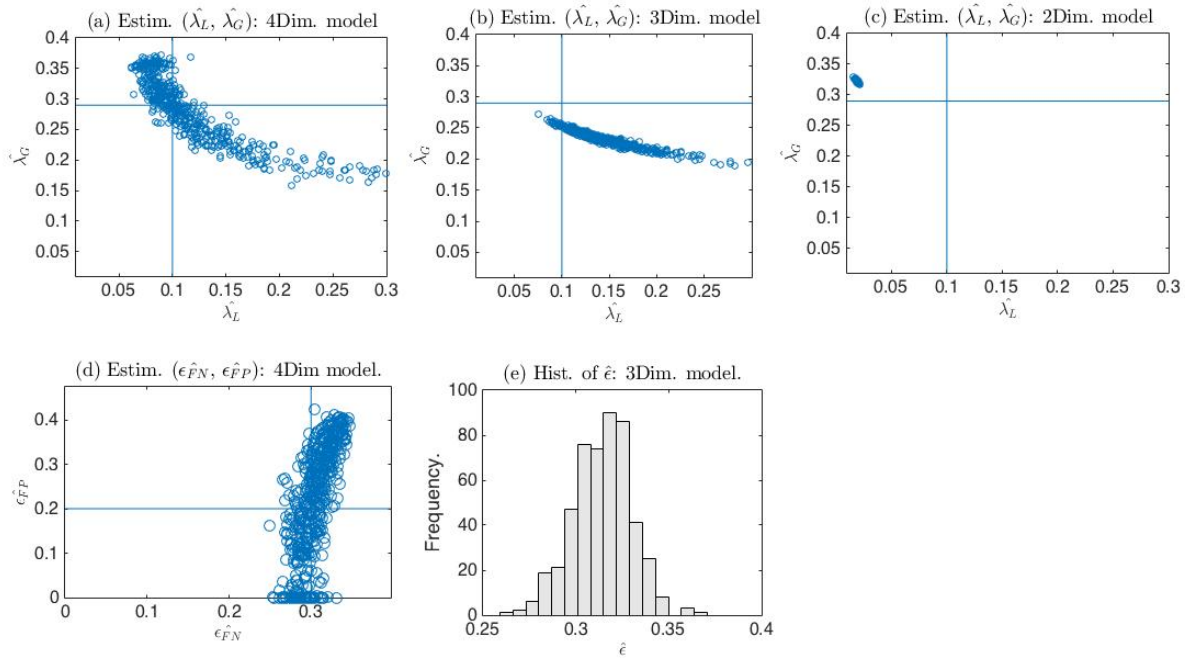

Figure 5.2: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$.

In figures 5.2 (b) and (c), the estimates of the two and three dimensional models are biased and imprecise when the misclassification probabilities are large and far apart from each other as theoretically expected.

5.4.4 Fitting the two, three and four dimensional models to the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2$.

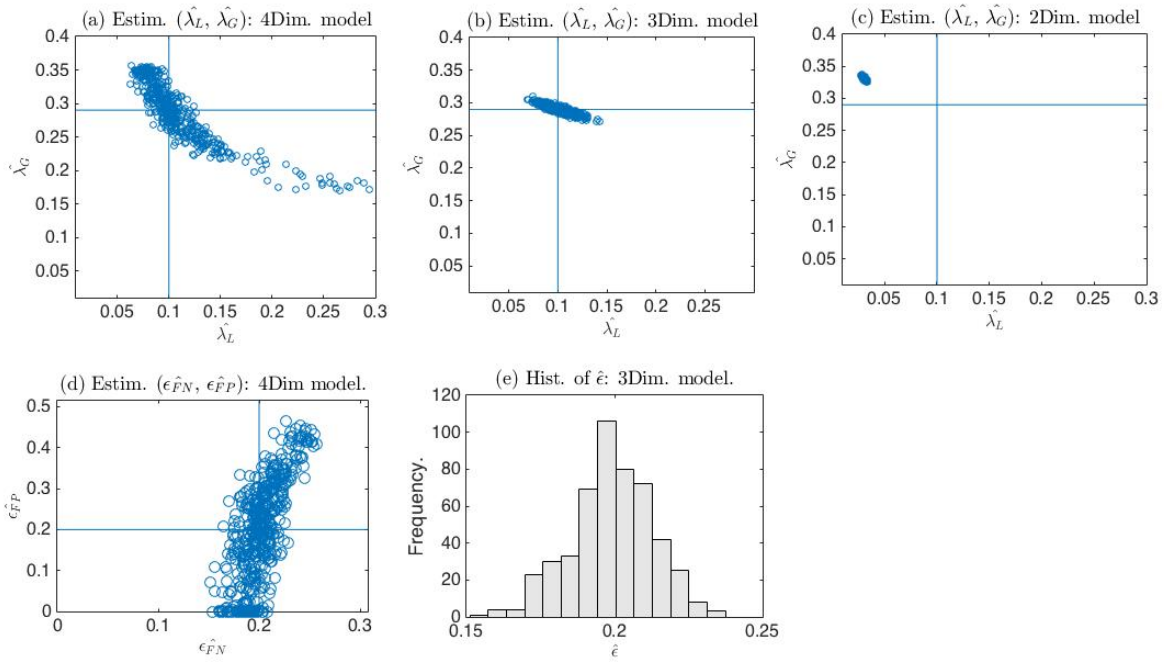

Figure 5.3: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2$.

In figures 5.3 (a) and (b), the scatter points of the estimates from the three and four dimensional models are centered at their true value with less variability for the three dimensional model, while those of the two dimensional model in 5.3 (c) are biased. The estimates of the three dimensional model are more precise than those of the two and four dimensional models.

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$			Theo.
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	Param.
$\hat{\lambda_L}$	0.084899	0.090811	0.10138	0.018974	0.14651	0.13265	0.03056	0.10032	0.117	0.1
$\hat{\lambda_G}$	0.3129	0.31015	0.28961	0.32242	0.23107	0.2744	0.33117	0.29025	0.28441	0.29
$\hat{\pi}$	0.38599	0.3865	0.4211	0.47438	0.52772	0.45322	0.42115	0.41985	0.4338	0.4199
$\hat{z}$	0.74206	0.74761	0.72958	0.56414	0.67592	0.71616	0.6369	0.72953	0.72469	0.7298
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.020239	N/A	N/A	0.30444	N/A	N/A	0.20185	N/A
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.097445	N/A	N/A	0.20979	N/A	N/A	0.19559	N/A
$\hat{\varepsilon}$	N/A	0.01074	N/A	N/A	0.31411	N/A	N/A	0.19921	N/A	N/A
$\hat{R_*}$	2.2495	2.2857	2.2164	1.5467	2.0074	2.1721	1.7365	2.2151	2.2004	2.2166

Table 5.1: Table of the mean of the parameter estimates of the three models.

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$		
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0015409	0.0044091	0.011434	0.00088506	0.056186	0.074851	0.0010531	0.012306	0.060019
$\hat{\lambda_G}$	0.0024536	0.0030933	0.017684	0.0018842	0.014262	0.05958	0.002174	0.0063406	0.047968
$\hat{\pi}$	0.0042913	0.0043876	0.029889	0.0031843	0.015945	0.10611	0.0035712	0.006532	0.084542
$\hat{z}$	0.0034378	0.0051624	0.015988	0.0024631	0.016208	0.057152	0.0027625	0.011531	0.044632
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.011379	N/A	N/A	0.019208	N/A	N/A	0.019286
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.06998	N/A	N/A	0.12818	N/A	N/A	0.12458
$\hat{\varepsilon}$	N/A	0.0072381	N/A	N/A	0.015834	N/A	N/A	0.01398	N/A
$\hat{R_*}$	0.016441	0.030185	0.064045	0.0050586	0.049663	0.23955	0.0075251	0.062329	0.18484

Table 5.2: Table of the standard deviation of the parameter estimates of the three models.

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$		
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.015179	0.01019	0.011506	0.081031	0.072919	0.081593	0.069448	0.012298	0.062322
$\hat{\lambda_G}$	0.023027	0.020387	0.017671	0.032475	0.060628	0.061531	0.041225	0.006339	0.048246
$\hat{\pi}$	0.034178	0.033682	0.029883	0.054578	0.10899	0.11112	0.0037814	0.0065256	0.085593
$\hat{z}$	0.012745	0.018553	0.015974	0.16567	0.056251	0.058699	0.092928	0.011522	0.044879
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.01137	N/A	N/A	0.019695	N/A	N/A	0.019355
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.069956	N/A	N/A	0.12842	N/A	N/A	0.12454
$\hat{\varepsilon}$	N/A	0.049788	N/A	N/A	0.066036	N/A	N/A	0.013988	N/A
$\hat{R_*}$	0.036808	0.075438	0.063981	0.66988	0.21497	0.24341	0.48018	0.062283	0.18536

Table 5.3: Table of the root mean square error of the parameter estimates of the three models.

5.5 Numerical simulations and inferences.

In this section, we studied the properties of the maximum likelihood estimates of the model parameters by exploring them along the diagonals of the misclassification probabilities permissible region $\{(\varepsilon_{FN}, \varepsilon_{FP}) : 0 \leq \varepsilon_{FN} \leq 0.2, 0 \leq \varepsilon_{FP} \leq 0.2\}$ in order to provide further insights into their behaviour.

We explored the estimates of the parameters along the diagonals of the misclassification probability region, $[0, 0.2]$, where $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, for $\varepsilon_{FP} \in [0, 0.2)$ using the following functions and subroutines described in the following.

Run the function FourThreeTwoDonFourfpos to simulate four dimensional final size epidemic data for misclassification probabilities, $\{(\varepsilon_{FN}, \varepsilon_{FP}) : 0 \leq \varepsilon_{FP} \leq \alpha\}$, $\varepsilon_{FN} = \alpha - \varepsilon_{FP}$, $\alpha < 0.5$ with Gamma(a, b) infectious period distribution. It then calculates other corresponding parameters of the three models from Gamma(a, b) infectious period distribution function and theoretical parameters λ_L, λ_G. It also calculates and plot the root mean square error of the estimates for misclassification probabilities $\varepsilon_{FP} \in [0, 0.5)$ with the subroutines in subsection 5.4.1.

While the function FourThreeTwoDonFourGraphSNsimhouses explores the estimates of the parameters along the vertical and horizontal axes of the misclassification Probabilities region using the following function and subroutines.

Run the function, FourThreeTwoDonFourGraphSNsimhouses to simulate four dimensional household epidemic data with Gamma(a, b) infectious period distribution function and msclassification probabilities ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$. It then explores the estimates of the three models along the vertical axis of the misclassification probabilities region with Gamma(a, b) infectious period distribution by holding ε_{FP} fixed while varying $\varepsilon_{FN} \in [0, 0.5)$. It also explores the estimates of the models along the horizontal by holding ε_{FN} fixed while varying $\varepsilon_{FP} \in [0, 0.5)$.

It computes and plot the root mean square error of the estimates for the three model using subroutines in subsection 5.4.1.

The estimation of the three models parameters employ similar subroutines with differences in the form of the function $q_i(n)$, where $q_i(n)$ is defined in equation (5.2.7). For the two dimen-

sional model, this function simply reduces to the final size probabilities, $q_i(n, \lambda_L, \pi, \varepsilon_{FN}, \varepsilon_{FP}) = P_i(n, \lambda_L, \pi)$, while for three dimensional model, it takes the form, $q_i(n, \lambda_L, \pi, \varepsilon_{FN}, \varepsilon_{FP}) = P_i(n, \lambda_L, \pi, \varepsilon)$, where we have assumed that $\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon$.

This is a special case of the four dimensional model in which the misclassification probabilities are the same. If $\varepsilon_{FN} = \varepsilon_{FP} = 0$, then the three and four dimensional models reduce to the two dimensional model, having the final size probabilities in equation (2.6.3). These models are nested in each other.

5.6 Comparison of the models.

We have estimated the parameters of the three models using the associated functions and subroutines discussed in section 5.5, along the diagonals of the two dimensional misclassification probabilities region, $[0, 0.5)$ with step size of 0.005 and presented results for misclassification probabilities in $[0, 0.2]$ owing to the repeated behaviour of the estimates in the remaining part of the misclassification region, $0.2 \leq \varepsilon_{FN} < 0.5$, $0.2 \leq \varepsilon_{FP} < 0.5$.

The mean, standard deviation, mean square error, root mean square error of the estimates are computed and the root mean square error are plotted in order to give insights on the fitness of the three models to the four dimensional final size epidemic data for theoretical parameters corresponding to small and large value of z away from its lower and upper boundaries.

These are accomplished by simulating household epidemic along the diagonal of the two dimensional misclassification probabilities region, $[0, 0.2]$ with theoretical parameters corresponding to $z = 0.4275$, given as $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $R_* = 1.4316$ and those corresponding to $z = 0.7298$ given as $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 04199$, $R_* = 2.2166$ respectively.

We then explored the estimates along the line, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, for each set of these parameter values. Where $\{\varepsilon_{FP} : 0 \leq \varepsilon_{FP} \leq 0.2\}$ with step size of 0.005.

We presented the plots of the root mean error of the estimates for the three models and tables of comparison, identifying regions where the estimates of the parameters of the models are precise on the four dimensional final size epidemic data.

5.6.1 Simulations with the theoretical parameter, $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.4316$.

We simulated household epidemic, with the following theoretical parameters, $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $R_* = 1.4316$ and misclassification probabilities, $\varepsilon_{\text{FN}} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{\text{FP}} \in [0, 0.2]$ with step size of $= 0.005$.

With theoretical parameters corresponding to $z = 0.42755$, we found the estimates of λ_L for the two dimensional model to be imprecise and biased especially when the misclassification probabilities increase from zero as in figure 5.4 (a). The two dimensional model is not a sufficient fit to the four dimensional final epidemic data. These behaviours can be observed for other parameters for the two dimensional model as in figures 5.4 (b)-(g).

The three dimensional model has precise estimates of λ_L for misclassification probability in $0.08 \leq \varepsilon_{FP} \leq 0.12$, while the four dimensional model is best if $0 \leq \varepsilon_{FP} \leq 0.08$ and $\varepsilon_{FP} \geq 0.17$. This shows that the four dimensional model has precise estimates of λ_L compared to those of the two and three dimensional models, if the misclassification probabilities are large and far apart from each other.

In the case of λ_G, the two dimensional model has imprecise and biased estimates, while those of the three dimensional model are precise if $0.08 \leq \varepsilon_{FP} \leq 0.01$, those of the four dimensional model are precise if, $0 \leq \varepsilon_{FP} \leq 0.075$ and $\varepsilon_{FP} \geq 0.115$.

In the case of π, the two dimensional model has precise estimates if, $0.02 \leq \varepsilon_{FP} \leq 0.025$, while the three dimensional model has precise estimates, if $0.03 \leq \varepsilon_{FP} \leq 0.105$, the four dimensional model is best if, $0 \leq \varepsilon_{FP} \leq 0.015$ and $\varepsilon_{FP} \geq 0.111$.

In the case of z, we found that the two dimensional model is best if, $0.085 \leq \varepsilon_{FP} \leq 0.095$, while the three dimensional model is best if $0.1 \leq \varepsilon_{FP} \leq 0.11$. The estimates of the four dimensional model are precise if, $0 \leq \varepsilon_{FP} \leq 0.08$ and $\varepsilon_{FP} \geq 0.115$.

In the case of the false positive misclassification probability estimates, the three dimensional model is best, if $0.09 \leq \varepsilon_{FP} \leq 0.115$, while the four dimensional model is best if $0 \leq \varepsilon_{FP} \leq 0.085$ and $\varepsilon_{FP} \geq 0.120$ respectively.

In the case of the false negative misclassification probability, the three dimensional model is best if, $0.09 \leq \varepsilon_{FP} \leq 0.115$, while the four dimensional model is best if, $0 \leq \varepsilon_{FP} \leq 0.085$

and $\varepsilon_{FP} \geq 0.120$.

Similarly in the case of the threshold parameter, the two dimensional model is best if $0.09 \leq \varepsilon_{FP} \leq 0.1$, the three dimensional model is best, if $0.1 \leq \varepsilon_{FP} \leq 0.105$, while the four dimensional model is best, if $0 \leq \varepsilon_{FP} \leq 0.085$ and $\varepsilon_{FP} \geq 0.110$.

In summary, we see in figures 5.4 (a)-(g) that the estimates from the four dimensional model are more precise than those from the two and three dimensional models when the misclassification probabilities are large and far apart from each other.

However if $\varepsilon_{FP} = 0.1$ then those of the three dimensional models are precise since the false negative misclassification probability, $\varepsilon_{FN} = 0.1$ reduces to the false positive misclassification probability, which is a particular case of the four dimensional model.

The estimates from the three dimensional model are precise if the two misclassification probabilities are close to each other while those of the two dimensional model are best if the misclassification probabilities are zero or close to it.

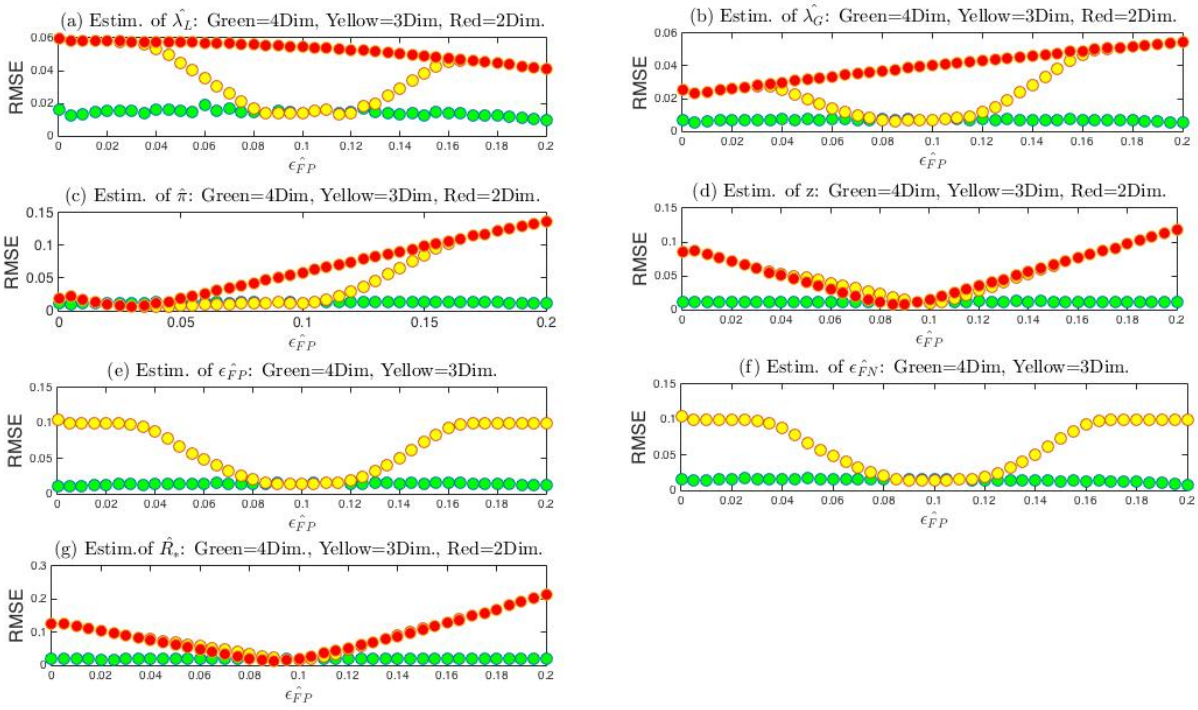

Figure 5.4: Plots of the root mean square error of the maximum likelihood estimates of the parameters for the three models when $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.4316$.

Figures 5.4 (a)-(g) are plots of the root mean square error of the maximum likelihood estimates of the parameters of the three models with regions of precision when the theoretical parameters corresponds $z = 0.4275$. We see that the root mean square error of the estimates from the four dimensional model are consistently stable throughout the misclassification probabilities region.

5.6.2 Simulations with theoretical parameters, $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$.

We simulated household epidemic with the following theoretical parameters along the line $\varepsilon_{\text{FN}} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{\text{FP}} \in [0, 0.2]$, step size $= 0.005$, $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$.

We then obtained the estimates of the parameters of the three models and presented

plots of their root mean square error in figures 5.5 (a)-(g) for a range of misclassification probabilities in $[0, 0.2]$.

From the simulation plots in figure 5.5 (a), we see that the estimates of λ_L from the two dimensional model are driven by bias and are precise if, $\varepsilon_{FP} \leq 0.02$, while the estimates of λ_L from three dimensional model are precise if, $0.050 \leq \varepsilon_{FP} \leq 0.165$. Those of the four dimensional model are precise if, $0 \leq \varepsilon_{FP} \leq 0.045$ and $\varepsilon_{FP} \geq 0.175$.

In the case of λ_G in figure 5.5 (b), the estimates of the two dimensional model are best if, $0 \leq 0 \leq \varepsilon_{FP} \leq 0.07$, those of the three dimensional model are best if, $0.075 \leq \varepsilon_{FP} \leq 0.145$, while those of the four dimensional model are best if $\varepsilon_{FP} \geq 0.150$.

Also, in the case of π in figure 5.5 (c), the estimates of the two dimensional are best if, $0.125 \leq \varepsilon_{FP} \leq 0.175$, those of the three dimensional model are best if, $0.07 \leq \varepsilon_{FP} \leq 0.120$, while those of the four dimensional model are best if, $0 \leq \varepsilon_{FP} \leq 0.065$, and $\varepsilon_{FP} \geq 0.18$.

In the case of z, the estimates of the two dimensional model are best if, $0.13 \leq \varepsilon_{FP} \leq 0.165$, those of the three dimensional model are best if, $0.065 \leq \varepsilon_{FP} \leq 0.125$, while those of the four dimensional model are best if, $0 \leq \varepsilon_{FP} \leq 0.06$, and $\varepsilon_{FP} \geq 0.17$.

In case of the false postive misclassification probability, ε_{FN}, the three dimensional model has precise estimates if, $0.05 \leq \varepsilon_{FP} \leq 0.165$, while the four dimensional has precise estimates if, $0 \leq \varepsilon_{FP} \leq 0.045$ and $\varepsilon_{FP} \geq 0.165$.

On the other hand, the estimates of the false negative misclassification probability from the three dimensional model are precise if $0.09 \leq \varepsilon_{FP} \leq 0.105$, while from the four dimensional model the estimates are precise if, $0 \leq \varepsilon_{FP} \leq 0.085$ and $\varepsilon_{FP} \geq 0.110$.

The threshold parameter, R_* has best estimates from the two dimensional model if, $0.14 \leq \varepsilon_{FP} \leq 0.165$, while it has best from the three dimensional model if, $0.065 \leq \varepsilon_{FP} \leq 0.135$. It has best estimates from the four dimensional model if, $0 \leq \varepsilon_{FP} \leq 0.060$ and $\varepsilon_{FP} \geq 0.170$.

In summary, with large misclassification probabilities, the estimates of the four dimensional model are more precise than those from the two and three dimensional models in agreement with the discussion in subsection 5.6.1.

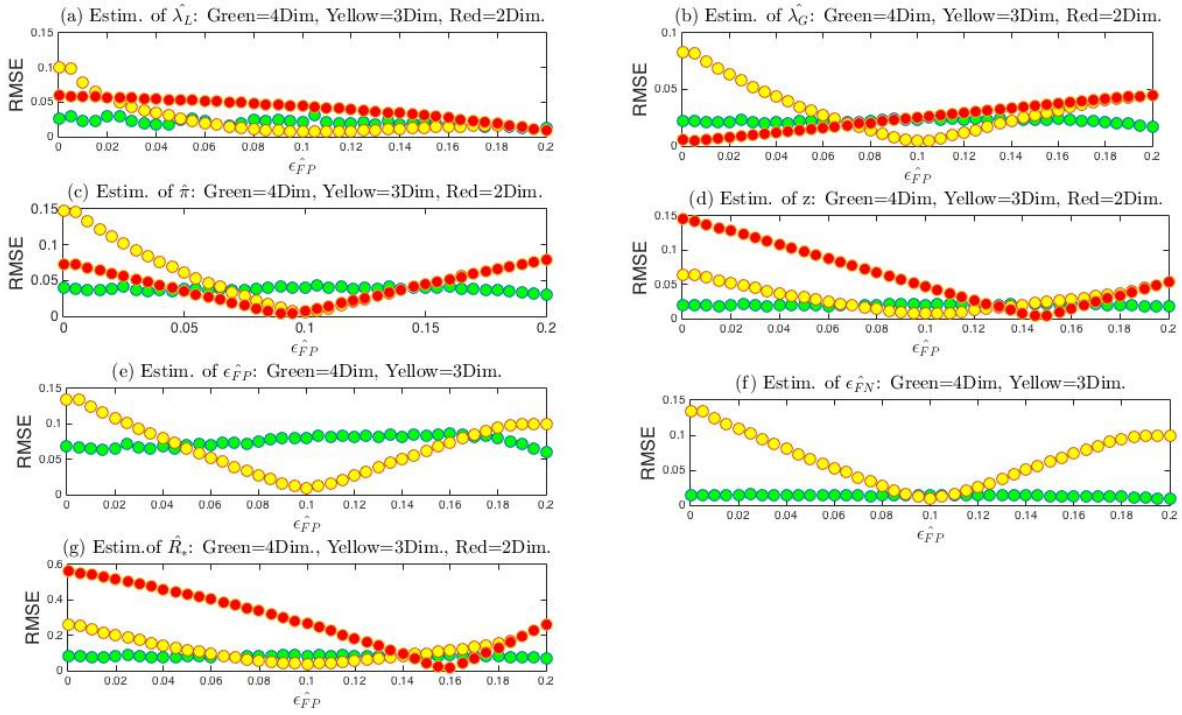

Figure 5.5: Plots of the root mean square error of the maximum likelihood estimates of the parameters for the three models when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$.

5.7 Summary of behaviour of the models.

From the statistical analyses of the models fitness to the four dimensional final size epidemic data, we see that precision of the estimates of the three models differs from parameter to parameter. For some parameters, the two dimensional model has precise estimate for ε_{FN}, ε_{FP} in the two dimensional misclassification parameter space, $\{(\varepsilon_{FN}, \varepsilon_{FP}) : \varepsilon_{FN} \in [0, 0.2], \varepsilon_{FP} \in [0, 0.2]\}$. While for some either the estimates of the three or four dimensional models are the best for misclassification probability in the permissible region.

However, figures 5.4 (a)-(g) and 5.5 (a)-(g) provide general summary of the properties of the estimates of the models on four dimensional final size epidemic data. Their behaviours along the diagonal of the misclassification probabilities region $[0, 0.2]$ are similar to those explored along the vertical and horizontal axes of the misclassification probabilities region $[0, 0.2]$ but have only chosen to present those of the former to avoid repetition.

5.8 Simulations and inferences of the three models.

Here, we studied the properties of the estimates of the three models on three dimensional epidemic data in the face of $\varepsilon \in [0, 0.5)$ using simulations with Gamma(a, b) infectious period distribution and pair of theoretical parameters (λ_L, λ_G) and the function, ThreefourTwoDimplotsEstimates with subroutines as follows.

1.) Run the function, ThreefourTwoDimplotsEstimates to simulate three dimensional final size epidemic data with Gamma(a, b) infectious period distribution, theoretical parameters λ_L, and λ_G, $\varepsilon \in [0, 0.5)$. It then calculates the other corresponding parameter of the three models and plot them. It also calculates the mean, standard deviation and root mean square error of the parameters estimates for the two, three and four dimensional models with Gamma(a, b) infectious period distribution and the following subroutines.

a.) LampaiD(mat), provides starting values for the two dimensional model parameters, λ_L and π according to [24].

b.) Enegloglik4(y, n, a, b, mat), computes the negative of the loglikelihood function associated with the three dimensional model with Gamma(a, b) infectious period distribution, the final size epidemic data and the starting parameters using inverse transformation of the parameter space.

c.) negloglik2(x, n, a, b, mat), computes the negative loglikelihood function associated with the two dimensional model using the parameters of Gamma(a, b) infectious period distribution, the final size epidemic data and the starting values of the parameter.

d) Misclass2(ε, n), computes the misclassification probabilites asssociated with the three dimensional model from the theoretical parameter ε and maximum household size n

e.) final_sizep$(a, b, \pi, n, \lambda_L)$ computes the final size probabilities associated with the two dimensional model from the parameters of Gamma(a, b) infectious period distribution, π, λ_L and maximum household size n.

f.) Misclass3$(a, b, n, \pi, \lambda_L, \varepsilon)$, computes the sum of the product of the misclassification probabilities and the final size probabilities associated with the three dimensional model.

g.) falseMisclass2$(\varepsilon_{FN}, \varepsilon_{FP}, n)$, computes the misclassification probabilities asssociated

with the four dimensional model.

h.) SIRfalsePmisclass$(a, b, n, \pi, \lambda_L, \text{fneg}, \text{fpos})$, computes the products of the misclassification probabilities and the final size probabilities associated with the loglikelihood function of the four dimensional model.

i.) pinf2$(a, b, \pi, \lambda_L \text{houses})$, calculates z and λ_G, where houses is the vector of household sizes from the parameters of Gamma(a, b) infectious period distribution, model parameters π, λ_L and vector of household sizes.

j.) RSTER2$(a, b, c, \lambda_L, \lambda_G, \text{houses})$ calculates the threshold parameter, R_* from Gamma(a, b) infectious period distribution, theoretical parameters λ_L, λ_G and vector of household sizes.

For a range of misclassification probabilities in the permissible region $\varepsilon \in [0, 0.5)$, we explored the estimates of the models with the function, TwoDonThreeSNsimhouses2 and subroutines as follows.

2. Run the function ThreefourTwoDimplotsRMSE, to simulate three dimensional final size epidemic data with Gamma(a, b) infectious period distribution, theoretical parameters λ_L, and λ_G, $\varepsilon \in [0, 0.5)$. It then calculates other parameters of the three models with Gamma(a, b) infectious period distribution, computes the mean, standard deviation and root mean square error of the parameters of the three models and plot their root mean square error of the estimates using the program function, ThreefourTwoDimplotsEstimates and subroutines as follows,

Using the program functions in (1) and (2), we present plots of the estimates, tables of mean, standard deviation and the root mean square error in section 5.9 and in section 5.10, we explored the estimates further with theoretical parameters away from their boundarires and then examined the precision of the estimates for $\varepsilon \in [0, 0.1]$.

Finally in section 5.11, we present table of summary of performance for the three models on final size epidemic data.

5.8.1 Fitting the two, three and four dimensional models to the three dimensional final size epidemic data.

Using the function, ThreefourTwoDimplotsEstimates we simulate three dimensional household epidemic with Gamma$(2, 2.05)$ infectious period distribution, theoretical parameters, $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$, misclassification probabilities, $\varepsilon \in [0.01, 0.0.2]$ and $\varepsilon = 0.2$, household structure in [1] but fifty times its population size, minimum epidemic size of 1000. We then estimate and plot the parameters of the three models as follows,

5.8.2 Fitting the two, three and four dimensional models to the three dimensional simulated final size epidemic data, when $\varepsilon = 0.01$.

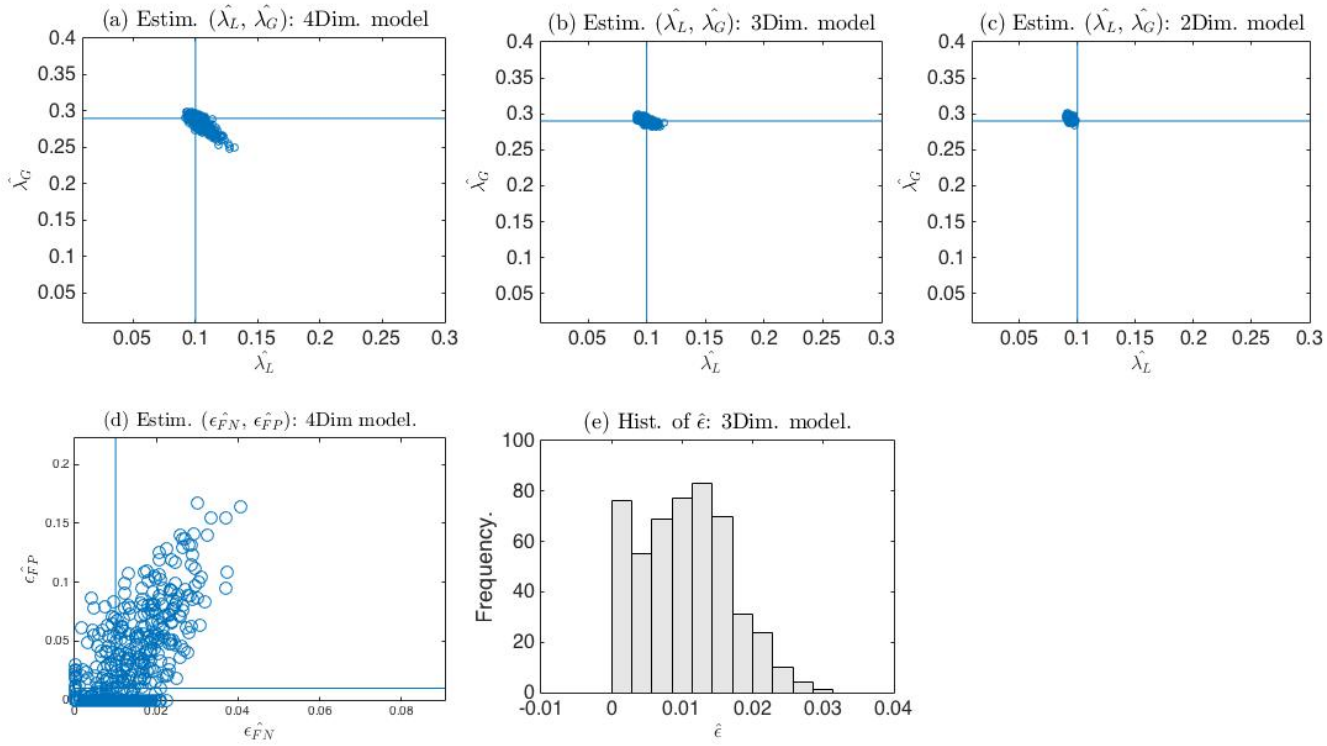

Figure 5.6: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.01$.

From figure 5.6, (c) the two dimensional models is beginning to struggle fitting to the three and four dimensional data when $\varepsilon = 0.01$, while those of the three and four dimensional

models are unbiased and precisely estimated as in figures 5.6 (a) and (b).

5.8.3 Fitting the two, three and four dimensional models to the three dimensional simulated final size epidemic data, when $\varepsilon = 0.02$.

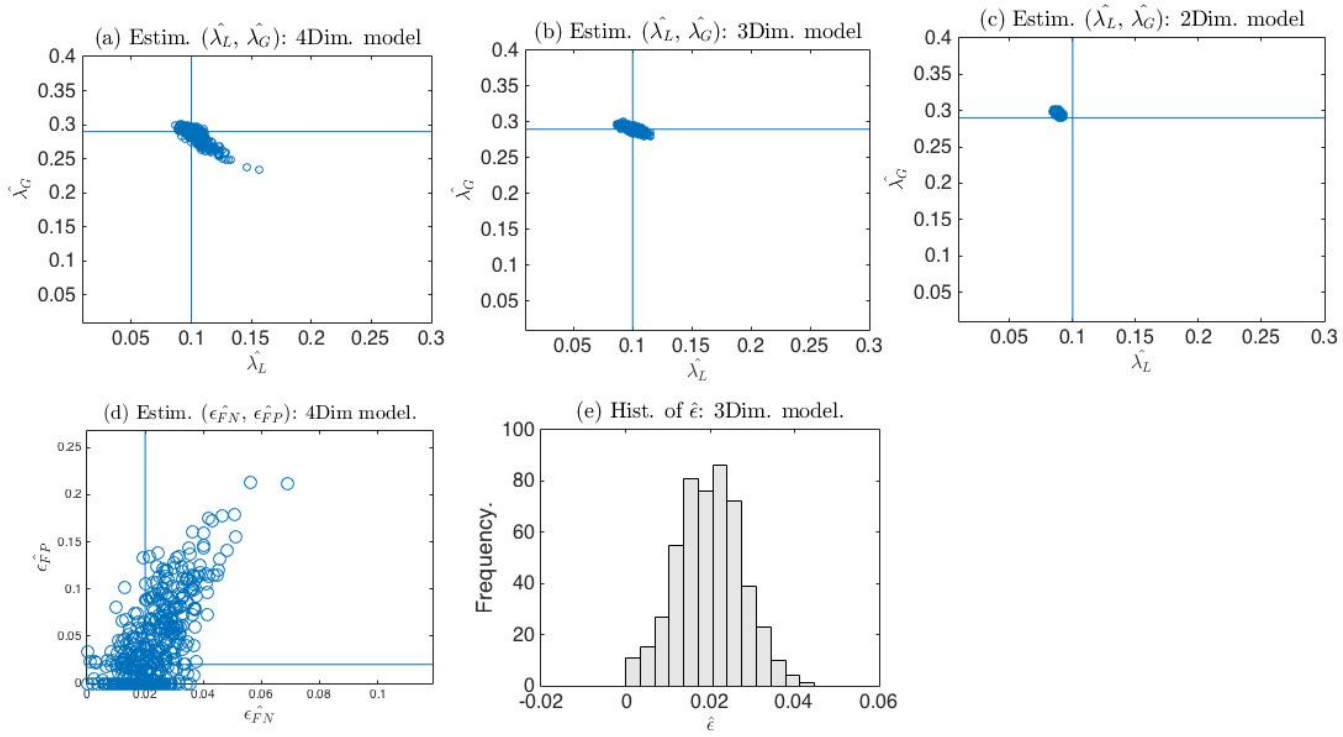

Figure 5.7: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.02$.

From figures 5.7 (a), (b) and (c), we see that when $\varepsilon = 0.02$, the parameter estimates from the two dimensional model become biased and imprecise, while those of the three and four dimensional models are unbiased and precise.

5.8.4 Fitting the two, three and four dimensional models to three dimensional simulated final size epidemic data, when $\varepsilon = 0.2$.

From figure 5.8 (c), we see that estimates from the two dimensional model are biased and imprecise while those from the three and four dimensional models in figures 5.8 (a) and (b)

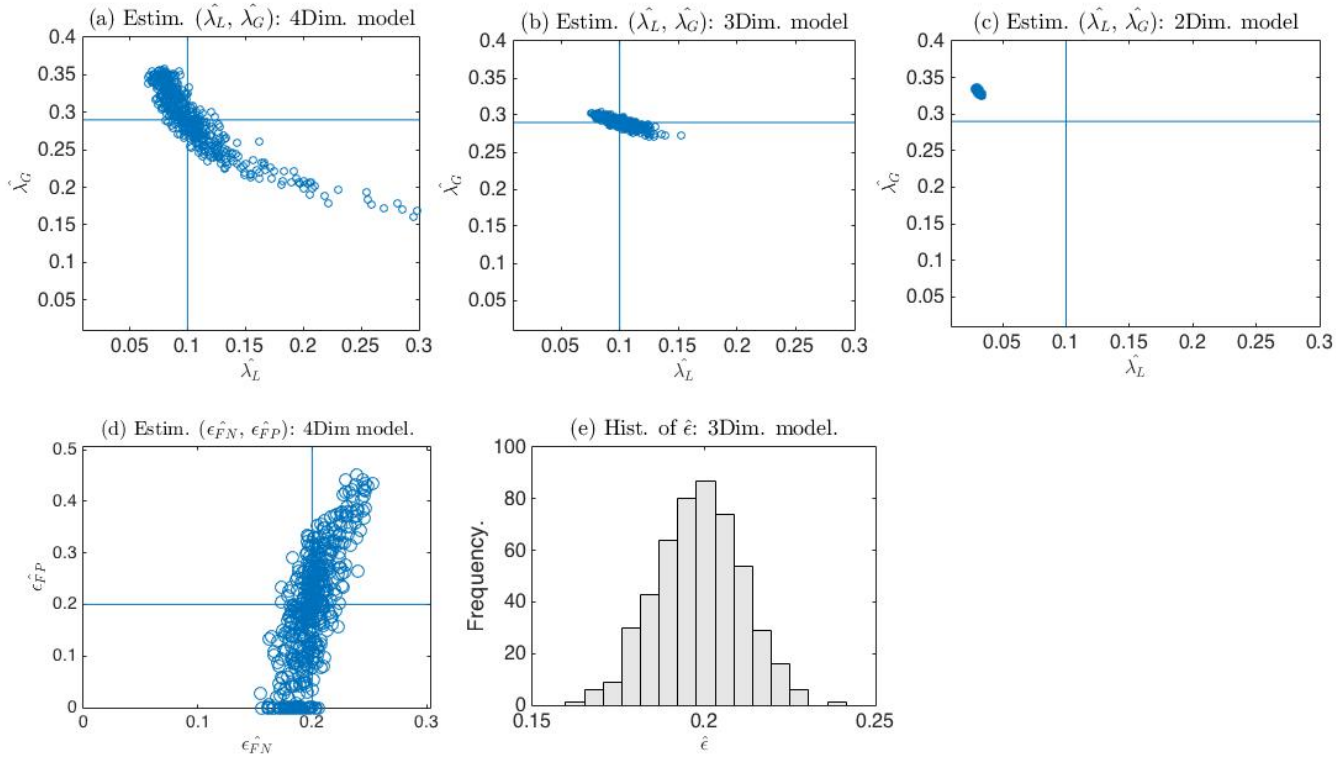

Figure 5.8: Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.2$

are precise and unbiased as expected.

Thus, with large misclassification probability $\varepsilon = 0.2$ the three and four dimensional models are the apropriate fit to three dimensional epidemic data. The three dimensional model with less number of parameters is often chosen in line with the principle of parsimony.

5.9 Table of mean, standard deviation and root mean square error of the estimates for the two, three and four dimensional models, when $\varepsilon = 0.01, 0.02$ and $\varepsilon = 0.2$.

We see from Table 5.4 that the maximum likelihood estimates of the two dimensional models are precise only when the misclassification probability is close to 0 and hence outperforms the three and four dimensional models, otherwise those of the three and four dimensional models have better precision.

In general, the three and four dimensional models outperforms the two dimensional model when the misclassification probability is far from 0.

Par.	Misclassification probability and model.									
	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	Theo Par.
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2	N/A
$\hat{\lambda_L}$	0.094069	0.10023	0.10309	0.088669	0.099995	0.10278	0.030611	0.099752	0.11092	0.1000
$\hat{\lambda_G}$	0.29291	0.28987	0.28513	0.29577	0.29005	0.28576	0.33108	0.29038	0.28831	0.29
$\hat{\pi}$	0.41882	0.42013	0.42827	0.41765	0.41995	0.42737	0.42125	0.41992	0.4268	0.4199
$\hat{z}$	0.72472	0.72974	0.72572	0.72004	0.72962	0.72606	0.6369	0.72901	0.72763	0.7298
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.013024	N/A	N/A	0.022364	N/A	N/A	0.20014	N/A
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.030605	N/A	N/A	0.037319	N/A	N/A	0.18729	N/A
$\hat{\varepsilon}$	N/A	0.010366	N/A	N/A	0.019881	N/A	N/A	0.19867	N/A	N/A
$\hat{R_*}$	2.188	2.2167	2.2018	2.161	2.2159	.2029	1.7367	2.2124	2.2105	2.2166

Table 5.4: Mean of the parameter estimates of the two, three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model.

Par.	Misclassification probability and model.								
	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.0014753	0.0042391	0.0066601	0.0014404	0.0050092	0.0081937	0.0010274	0.01126	0.047973
$\hat{\lambda_G}$	0.0024073	0.0031445	0.0090772	0.0023992	0.0032744	0.010904	0.0022198	0.0060073	0.044444
$\hat{\pi}$	0.0046921	0.0048705	0.015706	0.0046184	0.0048262	0.018881	0.0037478	0.0067691	0.077563
$\hat{z}$	0.0038545	0.0049281	0.0093312	0.0037818	0.0056181	0.011074	0.0029132	0.010631	0.040697
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.0079028	N/A	N/A	0.0091772	N/A	N/A	0.017635
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.037529	N/A	N/A	0.043826	N/A	N/A	0.11707
$\hat{\varepsilon}$	N/A	0.0064795	N/A	N/A	0.0077986	N/A	N/A	0.012364	N/A
$\hat{R_*}$	0.01685	0.024641	0.039134	0.016008	0.028788	0.046484	0.0076039	0.057059	0.16838

Table 5.5: Standard deviation of the parameter estimates of the two, three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model.

Par.	Misclassification probability and model.								
	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.006111	0.0042409	0.0073379	0.011422	0.0050042	0.0086443	0.069397	0.011252	0.049154
$\hat{\lambda_G}$	0.0037761	0.0031442	0.0073379	0.0062438	0.0032715	0.0086443	0.041142	0.0060135	0.049154
$\hat{\pi}$	0.0048073	0.0048714	0.017787	0.0051303	0.0048218	0.02029	0.0039827	0.0067624	0.077793
$\hat{z}$	0.0063716	0.0049235	0.010172	0.010457	0.0056151	0.011675	0.092936	0.010648	0.040714
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.0084544	N/A	N/A	0.0095477	N/A	N/A	0.017617
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.04278	N/A	N/A	0.0094679	N/A	N/A	0.11764
$\hat{\varepsilon}$	N/A	0.0064833	N/A	N/A	0.0077917	N/A	N/A	0.012423	N/A
$\hat{R_*}$	0.033136	0.024617	0.041799	0.057782	0.028766	0.048413	0.47997	0.057153	0.16832

Table 5.6: Root mean square error of the parameter estimates of the two, three and four dimensional models where, 2Dim=two dimensional model, 3Dim=three dimensional model and 4Dim=four dimensional model.

5.10 Simulations and inferences of the two and three dimensional models for $z \in [0, 1]$.

To enhance our understanding of the properties of the estimates in the face of misclassification probabilities in the permissible region, $[0, 0.5)$, we explored the estimates of the three models with two different sets of theoretical parameters with corresponding $z = 0.2144$ and $z = 0.7298$ away from their boundaries, simulation runs of 500, misclassification probabilities $\varepsilon \in [0, 0.1]$, with stepsize of 0.01, household structure in [1] and 50 times its population size, minimum

epidemic size of 1000, discussed in sections 4.2. We then simulate and estimate the models parameters, compute and plot the root mean square of the estimates using the function, ThreefourTwoDimplotsRMSE and subroutines in section 5.8.

Beginning with the theoretical parameters, $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\pi = 0.8999$, $z = 0.2144$, $R_* = 1.1653$, we simulate household epidemic, estimate the parameters of the models and examined their precision from the plots of the root mean square error for misclassification probabilities region $\varepsilon \in [0, 0.1)$.

5.10.1 Plots of the RMSE of the Parameter estimates when, $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\pi = 0.8999$, $z = 0.2144$, $R_* = 1.1653$.

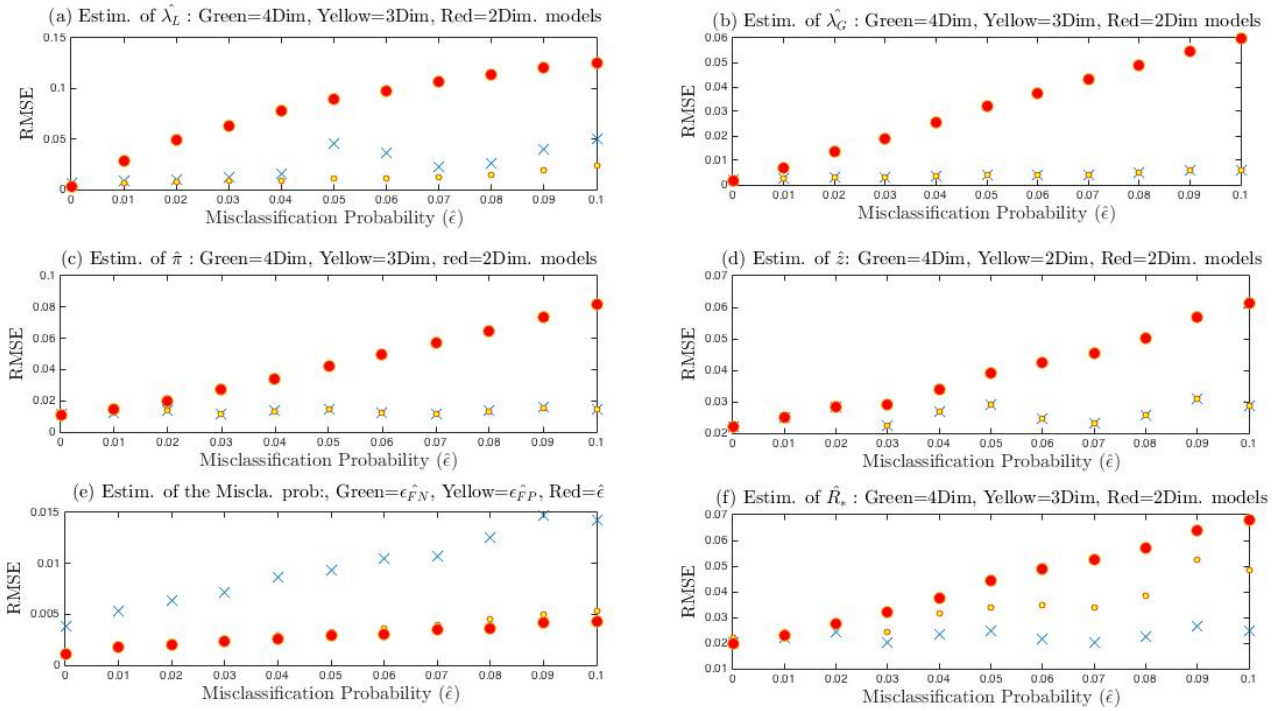

Figure 5.9: Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\pi = 0.8999$, $z = 0.2144$, $R_* = 1.1653$.

In figures 5.9 (a)-(f), we see that the estimates of the two dimensional model are precise compared to those of the three dimensional model if the misclassification probabilities are close to zero otherwise those of the three and four dimensional models are better.

5.10.2 Plots of the RMSE of the parameter estimates when $\lambda_L = 0.1,\ \lambda_G = 0.29,\ \pi = 0.4199,\ z = 0.7298,\ R_* = 2.2166.$

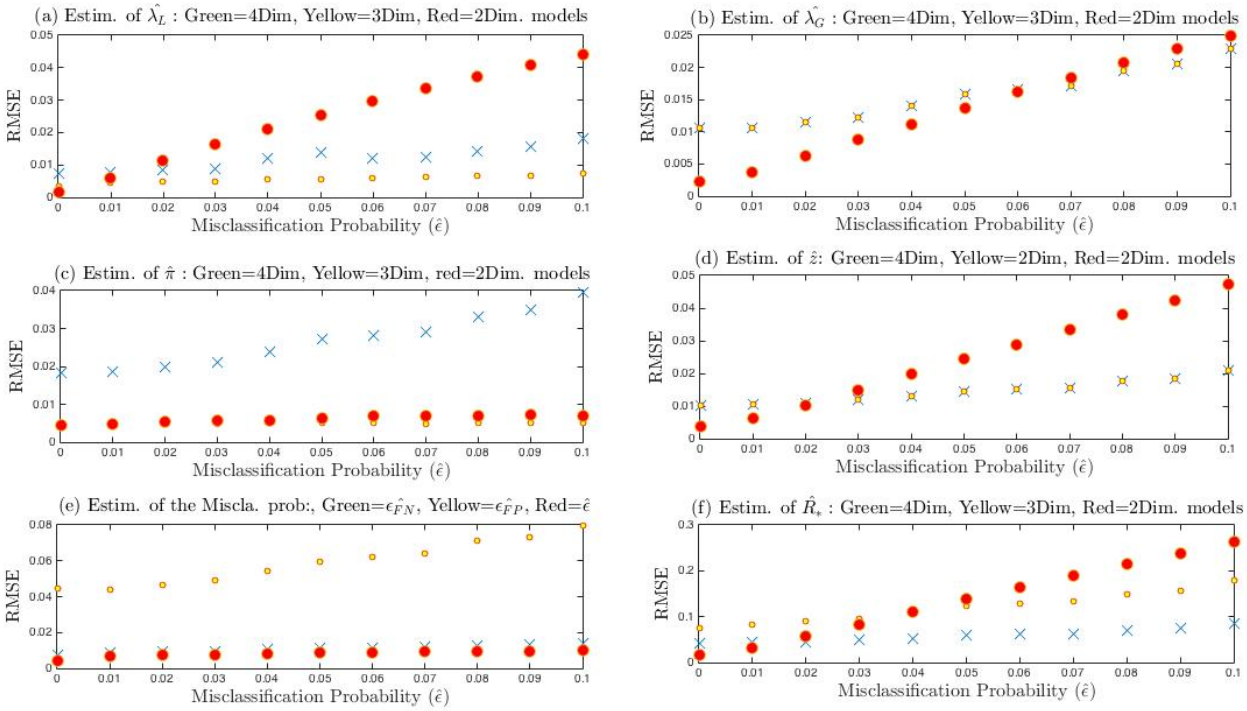

Figure 5.10: Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $\lambda_L = 0.1,\ \lambda_G = 0.29,\ \pi = 0.4199,\ z = 0.7298,\ R_* = 2.2166.$

In Figures 5.10 (a)-(f), similar pattern of behaviour are observed except that the estimates of λ_G in figure 5.10 (c) for the four dimensional are less precise than those of the two dimensional model. This may be attributable to the size of the proportion infected z as compared to its behaviour with $z = 0.2144$ in figure 5.9(c).

5.11 Summary of performance of the two, three and four dimensional models on final size epidemic data.

Here, we examined regions where the models outperform each other on the three dimensional final size household epidemic data for the set of theoretical parameters and misclassification probabilities $\varepsilon \in [0, 0.1]$. For example, the two dimensional model is found to be sufficient on the three dimensional final size epidemic data if ε is close to 0, while the three and four dimensional model are sufficient model fit if the misclassification probability is large. These properties are summarised in table 5.7.

The estimates of the two dimensional model are initialised according to [24] with minimum computational cost. For example from the [1] A(H3N2) Tercumseh Michigan epidemic the computational time for the estimates is 1.2 seconds, while those of the Seattle 1975-1976 B(H1N1) epidemic is 9 seconds, those of 1978-1979 A(H1N1) epidemic is 4.2 seconds.

In summary, the computational time required for convergence of the maximum likelihood estimates depends on the choice of the starting values and population size. With appropriate choice of the starting values away from the boundaries and large population size the computational time is large compared to small population size. However inadequate population size leads to lack of information and hence makes convergence of the estimates impossible.

Estimation Method.	Truth Simulated Data		
	Two Dimensional Simulated Data $\varepsilon = 0$, No Noise. Input Parameters, λ_L, λ_G	Three Dimensional Simulated Data $\varepsilon \neq 0$ Input Parameters, $\lambda_L, \lambda_G, \varepsilon$	Four Dimensional Simulated Data. $\varepsilon_{FN} \neq \varepsilon_{FP} \neq 0$ Input Parameters, $\lambda_L, \lambda_G, \varepsilon_{FN}, \varepsilon_{FP}$
Two Dimensional Optimisation Parameters Estimates, $\hat{\lambda_L}, \hat{\pi}, \varepsilon = 0$	Works well, with precise estimates for ε closer to 0, than three Dimensional Optimization Average computational time=1 seconds with Population size in [1]and [28] B(H1N1) epidemic data and initial estimates according [24] and 0.35 seconds for [28] Seattle A(H1N1)	Does not work well It give imprecise and biased estimates	Does not work well. It gives imprecise and biased estimates
Three Dimensional Optimisation Parameters Estimates, $\hat{\lambda_L}, \hat{\pi}, \hat{\varepsilon}, \varepsilon \neq 0.$	Works well but with less precision than two Dimensional Optimization.	Works well but with better precision $\forall \varepsilon \geq 0.005$. even for z close to the boundaries. Here λ_L and λ_G are initialised according to [24], while ε values are chosen away from its boundary as starting value. With the [1] and Seattle B(H1N1) epidemic data. the average computation time =9 seconds, while it is 0.42 seconds for the Seattle A(H1N1) epidemic data. Convergece of the estimates depends on the choice ε within the permissible region.	Works well only if the misclassification probabilities are close to each Otherwise does not work well
Four Dimensional Optimisation Parameter Estimates $\hat{\lambda_L}, \hat{\pi}, \varepsilon\hat{_{FN}}, \varepsilon\hat{_{FP}}.$ $\varepsilon_{FN} \neq \varepsilon_{FP} \neq 0$	Works well with less precision than two but better precision	Works well with approximately same precision as the three dimensional model if the misclassification probabilities are close	Works well with better precision $\forall \varepsilon_{FP}, \varepsilon_{FN} \in [0, 0.5)$ if the misclassification probabilities are far apart, where ε_{FN} and ε_{FP} are chosen away from the boundary of the misclassification probabilities permissible region and other parameters of the model are initialised according to [24] Convergence of the estimates depends on the choice of the starting values of ε_{FN} and ε_{FP}.

Table 5.7: Table of comparison of optimisations and models on the two, three and four dimensional simulated final size epidemic data.

Chapter 6

Chi-square goodness of fit test.

6.1 Introduction.

In this section, we fitted the three models to the final size epidemic data using the Pearson chi-square goodness of fit statistic with chi-square distribution function χ^2_v under the null hypothesis.

The Pearson chi-square goodness of fit test is meant to compare differences between the observed frequencies of the data with the expected frequencies, which are obtained according to a specific hypothesis. It compares the sample obtained with the hypothesized distribution to see if it fits the data. Here, v is the degrees of freedom of the test.

We also employed the Kolmogorov-Smirnov goodness of fit test to provide further insights on which sample data from the three models is from the hypothesized distribution.

These are accomplished by plotting the density histograms of the chi-square goodness of fit statistics for the three models each superimposed with their corresponding theoretical chi-square distribution functions. Their mean, variance are computed including the proportion of the simulations rejected from the Pearson chi-square goodness of fit test at the upper 5% point. Plots of the mean and variance of the Pearson chi-square goodness of fit statistics of the models are studied. Including the estimate of the models parameters, their Pearson chi-square goodness of fit statistics for $\varepsilon \in [0, 0.1]$, which corresponds to the three dimensional final size epidemic data and for ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$ corresponding to the four dimensional

epidemic data.

In the case of the four dimensional model, we explored the estimates along the diagonals (slicing through the two dimensional misclassification probabilities region) and along the vertical axes of the misclassification probabilities region for $\varepsilon_{FN}, \varepsilon_{FP} \in [0, 0.2)$. The Pearson chi-square goodness of fit statistics of the three models are then computed together with their mean and variance.

Plots of the mean and variance of the Pearson chi-square goodness of fit statistics are obtained including those of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test with the upper 5% point.

The empirical cumulative distribution function of the Pearson chi-square statistics of the three models together with the cumulative of the hypothesized chi-square distribution functions are also plotted.

This chapter is organised as follows:

In section 6.2, we discussed the degrees of freedom of the chi-square goodness of fit test with examples from [1] household epidemic data in section 6.3. In section 6.4, we discussed the likelihood ratio chi-squared goodness of fit test, while in section 6.5, we discussed the Kolmogorov-Smirnov goodness of fit test.

In section 6.8, we fitted the models to the two dimensional final size epidemic data and plotted the density histograms of the Pearson chi-square goodness of fit statistics. We also plotted the empirical cumulative distribution functions of the Pearson chi-squared statistics for the three models and those of cumulative of the chi-square distribution functions. We computed the mean and variance of the Pearson chi-square goodness of fit statistic.

In sections 6.9 and 6.10, we fitted the three models to the three dimensional final size epidemic data, plotted the density histograms of the Pearson chi-square statistics, those of the empirical cumulative distribution function of the Pearson chi-squared statistics and the cumulative of the hypothesized chi-square distribution functions for the three models.

We plotted the mean and variance of the Pearson chi-square goodness of fit statistics. for the three models for simulated epidemic data including also plot of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test with theoretical

parameters corresponding to $z = 0.2144$ and $z = 0.7298$ respectively over a range $\varepsilon \in [0, 0.1]$, with step size of 0.005.

In sections 6.11, 6.12 and 6.13, we fitted the models to the four dimensional final size epidemic data using the Pearson chi-square goodness of fit test and the Kolmogorov-Smirnov goodness of fit. We plotted the density histograms of the Pearson chi-square and those of the empirical cumulative distribution function of the Pearson chi-squared statistic superimposed with their theoretical chi-square distribution functions.

We also plotted the mean and variance of the Pearson chi-square goodness of statistic and the proportion of the simulations rejected from the Pearson chi-square goodness of fit test for theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$, over $\varepsilon \in [0, 0.2]$.

In section 6.14, we analysed and fitted the [1] Tecumseh Michigan epidemic data and in section 6.17 and discussed the properties of the models on the final size epidemic data.

6.2 Computation method of the Pearson chi-square goodness of fit statistic.

The expression $E_{i,j}$ is described as the expected number of j infectives from the household of size i when the null hypothesis is true and are computed from $q_j(i)$. Using [1] household structure in table 1.2, we present $E_{i,j}$ in table 6.1.

Household Size	Expected Number Infected in Household.					
	$E_{i,0}$	$E_{i,1}$	$E_{i,2}$	$E_{i,3}$	$E_{i,4}$	$E_{i,5}$
1	$N_1\hat{P}_0(1)$	$N_1\hat{P}_1(1)$	0	0	0	0
2	$N_2\hat{P}_0(2)$	$N_2\hat{P}_1(2)$	$N_2\hat{P}_2(2)$	0	0	0
3	$N_3\hat{P}_0(3)$	$N_3\hat{P}_1(3)$	$N_3\hat{P}_2(3)$	$N_3\hat{P}_3(3)$	0	0
4	$N_4\hat{P}_0(4)$	$N_4\hat{P}_1(4)$	$N_4\hat{P}_2(4)$	$N_4\hat{P}_3(4)$	$N_4\hat{P}_4(4)$	0
5	$N_5\hat{P}_0(5)$	$N_5\hat{P}_1(5)$	$N_5\hat{P}_2(5)$	$N_5\hat{P}_3(5)$	$N_5\hat{P}_4(5)$	$N_5\hat{P}_5(5)$

Table 6.1: Using [1] household structure and for $q_j(i)$, we have the expression for $E_{i,j}, i = 1, \ldots, n$, $j = 0, \ldots, i$. Where the final size probabilities are computed using the estimates of the model parameters for the corresponding dimensional model.

The Pearson chi-square statistic is given by,

$$\chi^2 = \sum_{i=1}^{maxh} \sum_{j=0}^{i} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{6.2.1}$$

where $O_{i,j}$ are the observed number of households of size i with j infectives in the final size epidemic data. These values can be identified from the [1] final size epidemic data in table 1.2. For example, $O_{1,0} = 110$, $O_{1,1} = 23$, $O_{2,0} = 149$, $O_{2,1} = 27$, $O_{2,2} = 13$ etc.

Each row of table 1.2 has one constraint, the total number of households of that size. The degrees of freedom of the test are obtained by subtracting one from the total number of cells, then summing them all together and subtracting the number of parameter estimates of the model, r from it.

If c_i is the total number of cells corresponding to the households of size i, then we can evaluate the degrees of freedom of the test as,

$$c_i - 1 = \sum_{j=0}^{i} c_{i,j} - 1,\ i = 1, 2 \ldots, \text{maxh}, \tag{6.2.2}$$

where maxh is the maximum household size, $c_{i,j}$ is the cell with j observed number of infectives in the household of size i defined by

$$c_{i,j} = \begin{cases} 1,\ if\ i = 1, 2, \ldots, maxh,\ j = 0, 1, \ldots, i \\ 0,\ otherwise. \end{cases}$$

Adding the degrees of freedom together $c - 1$ gives,

$$\begin{aligned} c - 1 &= \sum_{i=1}^{maxh} (c_i - 1) \\ &= \sum_{i=1}^{maxh} (\sum_{j=0}^{i} c_{i,j} - 1). \end{aligned} \tag{6.2.3}$$

Then the chi-square statistic, $\chi^2_{c-maxh-r}$ is said to have $c - maxh - r$ degrees of freedom. Where r is the number of parameters in the model and $c - maxh = {}^{maxh(maxh+1)}/_2$. For

example, χ^2_k, where $k = c - maxh - r$, has mean k and variance $2k$. These properties are further discussed in section 6.3, using [1] Tecumseh Michigan Influenza A(H3N2) epidemic data in table 1.2.

6.3 Degrees of freedom of the Pearson chi-square goodness of fit test.

From table 1.2 and equation (6.2.2), the first row corresponds to households of size one and has two nonempty cells. We subtract one from it $(2 - 1 = 1)$. Also the second row, which corresponds to households of size two and has 3 nonempty cells, we subtract one $(3 - 1 = 2)$. The third row has 4 nonempty cells, and corresponds to households of size three, we subtract one, $(4 - 1 = 3)$, the fourth row has 5 nonempty cells and corresponds to households of size four, we subtract one $(5 - 1 = 4)$ and finally the fifth row has 6 non empty cells and corresponds to households of size five, we subtract one $(6 - 1 = 5)$. Adding these values together as in equation (6.2.3) gives the total $c - 1 = 15$.

The degrees of freedom of the test depends on the number of parameters estimated in the model.

If we employed the two dimensional model, the number of parameters estimated is $r = 2$, namely λ_L and π and the degrees of freedom of the test is then

$c - 1 - r = 15 - 2 = 13$.

If the three dimensional model is employed, then the number of parameters estimated is $r = 3$ and the degrees of freedom of the test is $c - 1 - 3 = 15 - 3 = 12$.

Also, if it is the four dimensional model, the number of parameters estimated is $r = 4$, the degrees of freedom is $c - 1 - 4 = 15 - 4 = 11$.

6.4 Likelihood ratio chi-squared goodness of fit test.

A similar test to the Pearson chi-square goodness of fit test with the same degrees of freedom is the likelihood ratio chi-squared test proposed by [29]. The likelihood ratio chi-squared test

has the same asymptotic distribution as the Pearson chi-square goodness of fit statistic with the test statistic given by,

$$\chi^2 = \sum_{i=1}^{maxh} \sum_{j=0}^{i} O_{i,j} \left(\ln \frac{O_{i,j}}{E_{i,j}} \right). \tag{6.4.1}$$

Here $O_{i,j}$ and $E_{i,j}$ are the observed and expected frequencies of the final size data from the optimisations and ln is the natural logarithm.

These statistics may sometimes differ by large amount for some dataset, however the choice on which of the test to use depends on individual preference [29]. In this thesis, we found no such wide discrepancies between the two tests and therefore ignored it and presented the inference with the Pearson chi-square goodness of fit test.

6.5 Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov goodness of fit test is employed to verify whether a random sample is from a particular distribution, $F(x)$. It compares the hypothesized distribution function $F(x)$ under the null hypothesis and the empirical distribution function of the sample, $S(x)$ defined as the fraction of $X_i s$ that are less than or equal to x, where $-\infty < x < \infty$, are the sample data.

The test statistic T, is the vertical distance between the hypothesized distribution function $F(x)$ and the empirical distribution function of the sample $S(x)$ [29]. It have different representations depending on the type of hypothesis been tested, namely the two sided test, and the other two one sided tests.

The two sided test with the Null and alternative hypotheses, $H_0 : F(x) = S(x), H_1 : F(x) \neq S(x)$, has the test statistic, $T = \sup_x |F(x) - S(x)|$, while the one sided test with the Null and alternative hypotheses, $H_0 : F(x) \leq S(x), H_1 : F(x) > S(x)$ has the test statistic, $T^+ = \sup_x (F(x) - S(x))$. The one sided test with the Null hypotheses, $H_0 : F(x) \geq S(x), H_1 : F(x) < S(x)$ has the test statistic, $T^- = \sup_x (S(x) - F(x))$ [29].

We have adopted the notation of the alternative hypotheses for the tail hypotheses in the Mathworks documentations [46] given as, 0, for the alternative of the two sided hypothesis test, $H_1 : F(x) \neq S(x)$, 1, for the alternative of the one sided hypothesis test, $H_1 : F(x) > S(x)$

and -1, for the alternative of the one sided hypothesis test, $H_1 : F(x) < S(x)$ respectively.

The function, kstest(Dataset, CDF, Alpha, Tail) in the Mathworks documentation [46] is employed to compute the critical value of the test using approximate formula or by interpolation in a table [46], valid for the range of $0.01 \leq \alpha \leq 0.2$ for the two-sided test and $0.05 \leq \alpha \leq 0.1$ for the one-sided tests respectively. Here, CDF is a two column matrix, having as its first column the sample data, the second column is made of the cumulative distribution function of the hypothesized distribution, Alpha is the chosen level of significance, while the tail represents the alternative hypotheses of the two and the one-sided hypotheses been tested.

Since the critical value is approximate, comparing it with the test statistic will give a different decision [46] and hence the comparison adopted is such that if there is good agreement between the empirical distribution $S(x)$ and hypothesised distribution $F(x)$, the P-values will be large compared to the level of significance of the test and the null hypothesis is then accepted. Small value of p cast doubt on the validity of the test [46]. We have employed these procedures to test the three models fitness to the final size epidemic data, at the default upper 5% level of significance with the decision rules, $h = 1$ and $h = 0$ for rejecting and not rejecting the null hypothesis respectively.

6.6 Proportion of the simulations rejected from the Pearson chi-square goodness of fit test.

Having computed the Pearson chi-square statistics corresponding to the three models as discussed in sections 6.7 and 6.3 respectively, we can test the null hypothesis at the default upper $\alpha = 5\%$ significance and reject the models or not depending on the value of the Pearson chi-squares statistics at the $1 - \alpha$ quantile of the chi-square distribution.

In a simulation experiment, we can also compute the proportion rejected which is the numerical approximation of the power function. It tells us how often we reject the null hypothesis when it is false. For example, if the two dimensional model is sufficient on the two dimensional final size epidemic data then we expect that $X_2 \approx \chi^2_{13}$, where X_2 is the Pearson

chi-square goodness of fit statistic from the two dimensional final size data and we do not reject the two dimensional model. If the two dimensional model is not sufficient on the two dimensional final size epidemic data then $X_2 \gg \chi^2_{13}$.

Similarly, if the three dimensional model is sufficient on the three dimensional final size epidemic data then we expect $X_3 \approx \chi^2_{12}$, where X_3 is the Pearson chi-square goodness of fit statistic from the three dimensional final size epidemic data and we do not reject the three dimensional model. If the three dimensional model is not sufficient on the three dimensional final size epidemic data then $X_3 \gg \chi^2_{12}$.

Also if the four dimensional model is sufficient on the four dimensional model then $X_4 \approx \chi^2_{11}$, where X_4 is the Pearson chi-square goodness of fit statistic from the four dimensional final size epidemic data and we do not reject the four dimensional model. If the four dimensional model is not sufficient on the four dimensional final size epidemic data then $X_4 \gg \chi^2_{11}$.

Using the upper $\alpha = 5\%$ points which corresponds to the $1-\alpha$ quantiles of the chi-square distribution given as 22.36, 21.03 and 19.68 respectively, these scenarios are better understood as follows.

We reject the two dimensional model, if $X_2 > 22.36$ when the two dimensional model is true. Also we reject the three dimensional model, if $X_3 > 21.03$, when the three dimensional model is true. In the same way, we reject the four dimensional model, if $X_4 > 19.68$ when the four dimensional model is true.

We compute the proportions of the simulation rejected at α level of significance obtained as follows.

If the Pearson chi-square statistic from the two dimensional model are $X_2 = 47, 12, 53, 57, 31$ then using the upper 5% point we sum the number of the simulations rejected as, sum($X >$ 22.36) and determine the proportion rejected or the power of the test as, sum($X_2 > 22.36$)/length(X), which in this case is 0.800.

If the model fits well to the final size epidemic data, then we expect the proportion rejected or the power of the test to be close to 0.05, while if it doesn't fit well, then we expect the proportion rejected or the power function to be close to 1.

Using this approach with $\varepsilon = 0, 0.1, 0.3$, we presented in figure 6.5 the proportion infected

for simulated epidemic data over a range of $\varepsilon \in [0, 0.1]$ with step size of 0.005 for the three models.

We have also extended this method to the four dimensional epidemic data in table 6.17 for a range of misclassification probabilities in $[0, 0.2]$, in table 6.9 and presented plots of the proportion rejected for the three models in figure 6.13 for misclassification probabilities in $\varepsilon_{FP} \in [0, 0.2]$ and theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$ respectively.

In general, the usual behaviour of the models in the face of varying misclassification probabilities in the permissible region $[0, 0.5)$ are observed. We found that without misclassification probabilities in the final size epidemic data, the proportion of the simulations rejected for the two dimensional model in 7.2 is small but with increasing misclassification probabilities the proportion of the simulations rejected for the two dimensional model increases towards 1 as theoretically expected. While the three dimensional model has a small proportion rejected when $\varepsilon = 0.1$ but with increasing misclassification probabilities towards its upper boundary, the four dimensional model has small proportion of the simulations rejected compared to the two and three dimensional model. This demonstrates the strength of the four dimensional model over the two and three dimensional models when the misclassification is large.

6.7 Pearson chi-square goodness of fit test on two dimensional final size epidemic data.

We fitted the three models to the two dimensional final size epidemic data as follows.

Run the function, TwoThreeandfourandTwoSNsimhouseschsqlik to simulate two dimensional final size household epidemic data having Gamma(a, b) infectious period distribution, with the theoretical parameters, λ_L and λ_G. It then calculates the parameters of the three models with Gamma(a, b) infectious period distribution, computes their pearson chi-square square statistics and plot their density histogram superimposed with their theoretical chi-square distribution. It also computes the mean and variance of the Pearson chi-square statistic from the three models.

It computes the proportion of the simulations rejected from the Pearson chi-square goodness of fit test, the empirical cumulative distribution from the three models and plot the empirical cumulative distribution function with the cumulative of the hypothesized chi-square square distribution function for the three models using the following subroutines.

a.) LampaiD(mat), provides starting values for the two dimensional model parameters, λ_L and π according to [24].

b.) Enegloglik4(y, n, a, b, mat), computes the negative of the loglikelihood function associated with the three dimensional model with Gamma(a, b) infectious period distribution, the final size epidemic data and the starting parameters.

c.) negloglik2(x, n, a, b, mat), computes the negative loglikelihood function associated with the two dimensional model with Gamma(a, b) infectious period distribution, the final size epidemic data and the starting values of the parameter.

d) Misclass2(ε, n), computes the misclassification Probabilities asssociated with the three dimensional model from the misclassification probability parameter ε and maximum household size n.

e.) final_sizep(a, b, π, n, λ_L) computes the final size probabilities from the parameters associated with the two dimensional model using Gamma(a, b) infectious period distribution, estimates of π, λ_L and maximum household size n. It also computes the Pearson chi-square statistics from the two dimensional model.

f.) Misclass3($a, b, n, \pi, \lambda_L, \varepsilon$), computes the sum of the product of the misclassification probabilities and the final size probabilities associated with the three dimensional model for the computation of the negative loglikelihood function. It also computes the Pearson chi-square statistic from the three dimensional model.

g.) falseMisclass2(ε_{FN}, ε_{FP}, n), computes the misclassification probabilities asssociated with the four dimensional model.

h.) SIRfalsePmisclass(a, b, n, π, λ_L, fneg, fpos), computes the sum of the products of the misclassification probabilities and the final size probabilities associated with the four dimensional model and the Pearson chi-square statistics of three models.

i.) pinf2(a, b, π, λ_Lhouses), calculates z and λ_G, from the parameters of Gamma(a, b)

infectious period distribution, model parameters π, λ_L and vector of household sizes.

j.) RSTER2$(a, b, c, \lambda_L, \lambda_G, houses)$ calculates the threshold parameter, R_* from the parameters of Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G and vector of household sizes, houses.

The expected frequencies are computed as $E_{i,j} = \hat{p}_i(j)N_i$, where $\hat{p}_i(j)$ are the final size probabilities and computed from the subroutine,

final_sizep(a, b, π, λ_L), N_i is the number of households of size $i = 1, 2, \ldots, n$, $j = 0, 1, \ldots, i$.

6.8 Numerical simulations on two dimensional final size epidemic data.

Using the procedures in 6.7, we simulate household epidemics with Gamma(a, b) infectious period distribution, theoretical parameters, $\lambda_L = 0.1$, $\lambda_G = 0.29, \pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$, minimum epidemic size of 1000, fifty times household size in [1], simulation runs of 1000 and plotted the density histograms of the Pearson chi-square goodness of fit statistic superimposed with their theoretical chi-square distribution. Also, we plotted the empirical cumulative distribution functions superimposed with the cumulative of the hypothesized chi-squared distribution function. The mean and variance of the Pearson chi-square goodness of fit statistic and the proportion of the simulations rejected from the Pearson chi-square goodness of fit test with the upper 5% points are also computed.

6.8.1 The Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests on two dimensional final size epidemic data.

From figures 6.1 (a), (c), (e), we found that the three models are not good enough on the two dimensional data. This clarity can be seen from plots of the empirical cumulative distribution function and hypothesized cumulative distribution in figures 6.1 (b), (d) and (f).

In this case the model with the smallest number of parameters is often chosen in line with the principle of parsimony.

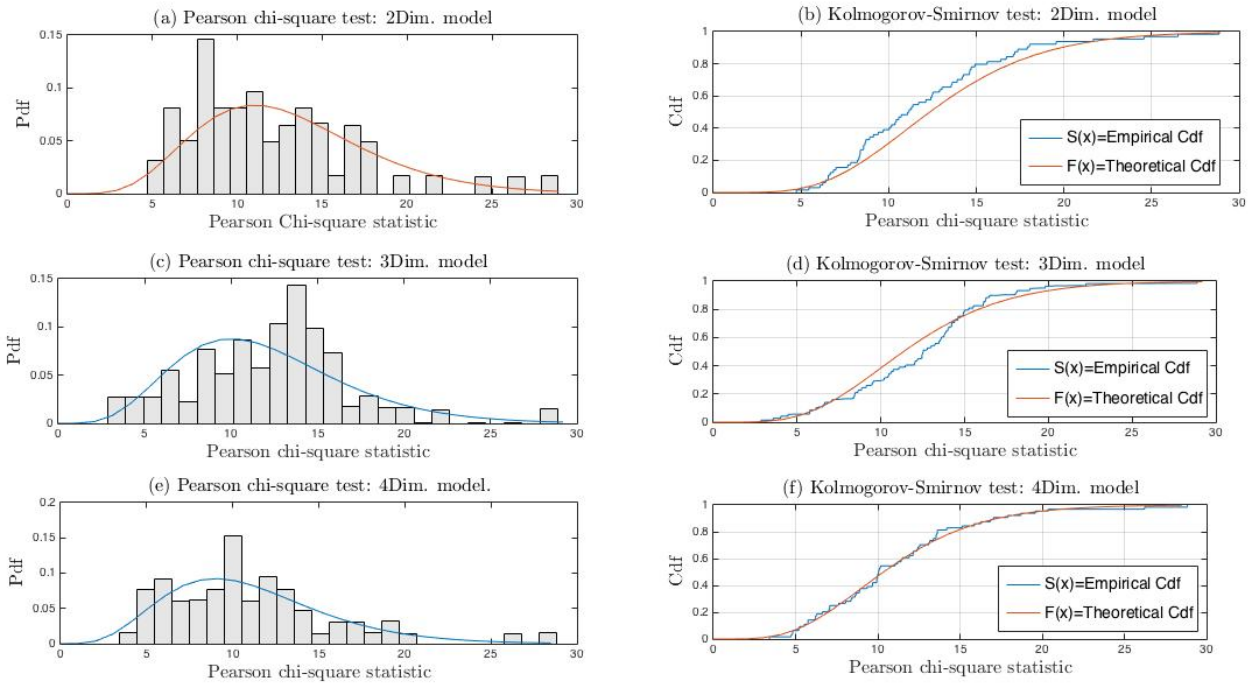

Figure 6.1: Density histograms of the Pearson chi-square goodness of fit and the Kolmogorov-Smirnov goodness of fit tests on the two dimensional final size epidemic data.

6.8.2 Table of mean and variance of the Pearson chi-square test on the two dimensional final size epidemic data.

Statistic	Two Dim. Model		Three Dim. Model		Four Dim. Model.	
	Sim. chi value	Theo. value	Sim. chi. value	Theo. value	Sim. chi. value	Theo. value
Mean	13.308	13	12.823	12	11.932	11
Variance	28.952	26	28.099	24	26.044	22

Table 6.2: Table of the mean and variance of the Pearson chi-square statistic of the models to two dimensional final size epidemic data. Where Sim. is the simulated values, chi. is the Pearson chi-square goodness of fit statistic, Theo. is the Theoretical mean and variance of the chi-square statistic.

The Mean and variance of the Pearson chi-square statistic for the three models, defined

here as the simulated mean and variance, are all approximately equal to their theoretical values in table 6.2 and are therefore approximately equal to the theoretical chi-square distribution function.

The two dimensional model is the simplest of the three models and therefore the preferred model fit to the two dimensional final size epidemic data as seen from figures 6.1 (a) and (b). Also from table 6.3 we see that the proportion of the simulations rejected is not close to 1, for the three models, as this will mean their misfit to the final size epidemic data. This provides further evidence that the three models fit fairly well to the two dimensional final size epidemic data.

Pearson chi-square statistic.	Upper 5% point	Proportion Rejected
χ^2_{13}	22.36	0.0940
χ^2_{12}	21.03	0.0960
χ^2_{11}	19.68	0.1400

Table 6.3: Table showing the proportion of the simulations rejected from the two dimensional final size epidemic data in figures 6.1 (a), (c) and (e).

From table 6.4, the null hypothesis from the two sided test is rejected from the two dimensional model owing to significant discrepancy between the empirical cumulative distribution function and the cumulative of the chi-square distribution function in one direction, while

Model	Tail (F>S F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.0000010 T=0.109964	h=0, p=0.220801 T= 0.038535	**h=1**, p=0.000005 T=0.109964
3Dim.	**h=1**, p=0.00008, T=0.100172	**h=1**, p=0.000040 T=0.100172	**h=1**, p=0.00100 T=0.082736
4Dim.	**h=1**, p=0.000029 T=0.105015	**h=1**, p=0.000144 T=0.093661	**h** = 1, p=0.000015, T=0.105015

Table 6.4: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the two dimensional final size epidemic data in figure 6.1 (b), (d), (f).

those of the three and four dimensional models were rejected because of the significance discrepancies in both directions.

6.9 The Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests on the three dimensional final size epidemic data.

Here, we explored the the parameter estimates of the three models with $\varepsilon = 0.0, 0.1, 0.3$ using the following function and subroutines.

Run the function ThreeandTwoDimoptonThreesimhousesChsqlik to simulate three dimensional epidemic data having Gamma(a, b) infectious period distribution, for $\varepsilon \in [0, 0.5)$ and theoretical parameters, λ_L and λ_G. It then calculates the parameters of the three models with Gamma(a, b) infectious period distribution. It computes the Pearson chi-square square statistic its mean and variance, the empirical cumulative distribution of the chi-square statistics of the three models and the proportion of the simulations rejected from the chi-square goodness of fit statistic at 5% significance. These are accomplished with the subroutines in subsection 6.7.

We extended our studies for a range of misclassification probabilities, $\varepsilon \in [0, 0.1]$ with step size of 0.005 for theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$ presented in section 6.10. These are achieved using the following function and subroutines.

Run the function, ThreefourTwoDimplotschsqlik to simulate household epidemic with Gamma(a, b) infectious period distribution for a range of $\varepsilon \in [0, 0.5)$, theoretical parameters, λ_L and λ_G. It then calculates the other parameters of the three models with Gamma(a, b) infectious period distribution over $\varepsilon \in [0, 0.5)$ and computes the Pearson chi-square goodness of fit and the chi-square difference goodnes of fit statistics of the three models. It plots their mean and variance including the proportion of the simulation rejected by the Pearson chi-square and chi-square difference test at 5% significance using the subroutines in subsection 6.7.

6.9.1 When the misclassification probability $\varepsilon = 0.1$.

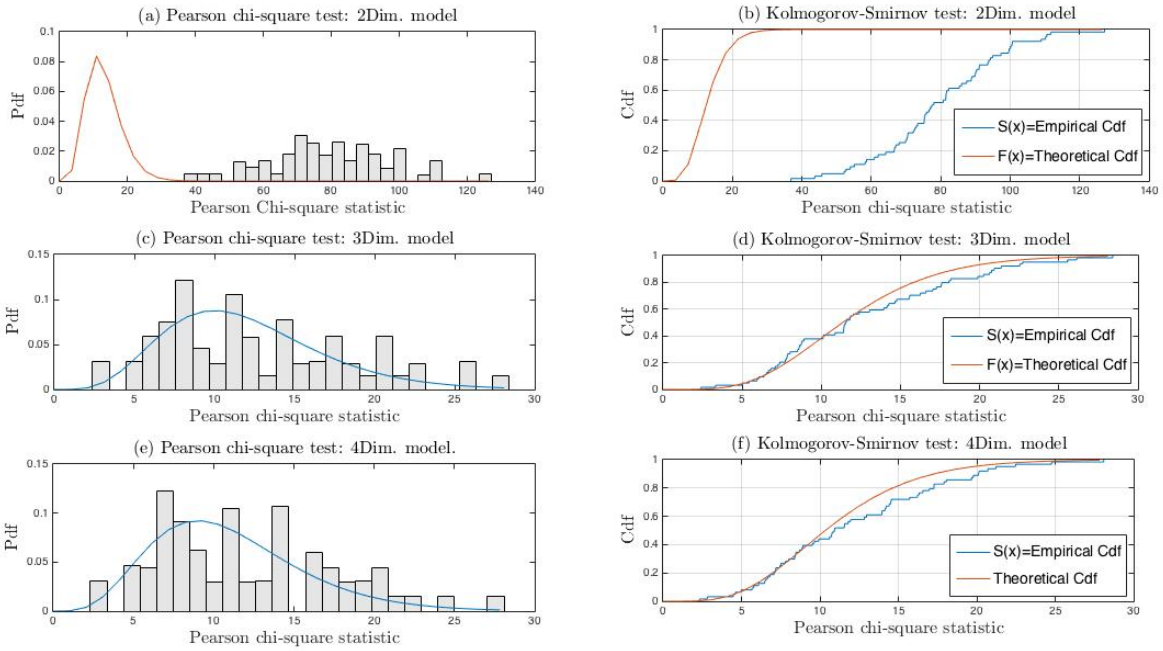

Figure 6.2: Density histograms of the Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests of the three models on the three dimensional final size epidemic data when $\varepsilon = 0.1$.

In figure 6.2 (a), we see that when the misclassification probability is large away from the lower boundary of the misclassification probability permissible region $[0, 0.5)$, the two dimensional model struggled fitting the final size epidemic data as shown by the density histograms and the Kolmogorov-Smirnov goodness of fit test. The three and four dimensional models sufficiently fit the three dimensional final size epidemic data.

From table 6.5, we see that with misclassification probability $\varepsilon = 0.1$, the null hypothesis is rejected for the two sided test at 0.05 significance level from the three models because of the significant discrepancies between the empirical cumulative distribution functions and the hypothesized chi-square distribution functions in one direction. However, the empirical cumulative distributions from the three and four dimensional models are better approximations of the cumulative of the chi-square distribution.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.0000 T=1.0000	h=0, p=1.00000, T=0.0000	**h=1**, p=0.0000, T=1.0000
3Dim.	**h=1**,p=0.03691, T=0.078971	h=0, p=0.106755, T=0.046966	**h=1**, p=0.001845, T=0.078971
4Dim.	**h=1**, p=0.00004, T=0.114159	h=0, p=0.133589, T=0.044534	**h=1**, p=0.00002, T=0.114159

Table 6.5: Summary of the Kolmogorov-Smirnov test for the upper 5% points for the three dimensional final size epidemic data when $\varepsilon = 0.1$ in figures 6.2(b), (d) and (f).

6.9.2 When the misclassification probability $\varepsilon = 0.3$.

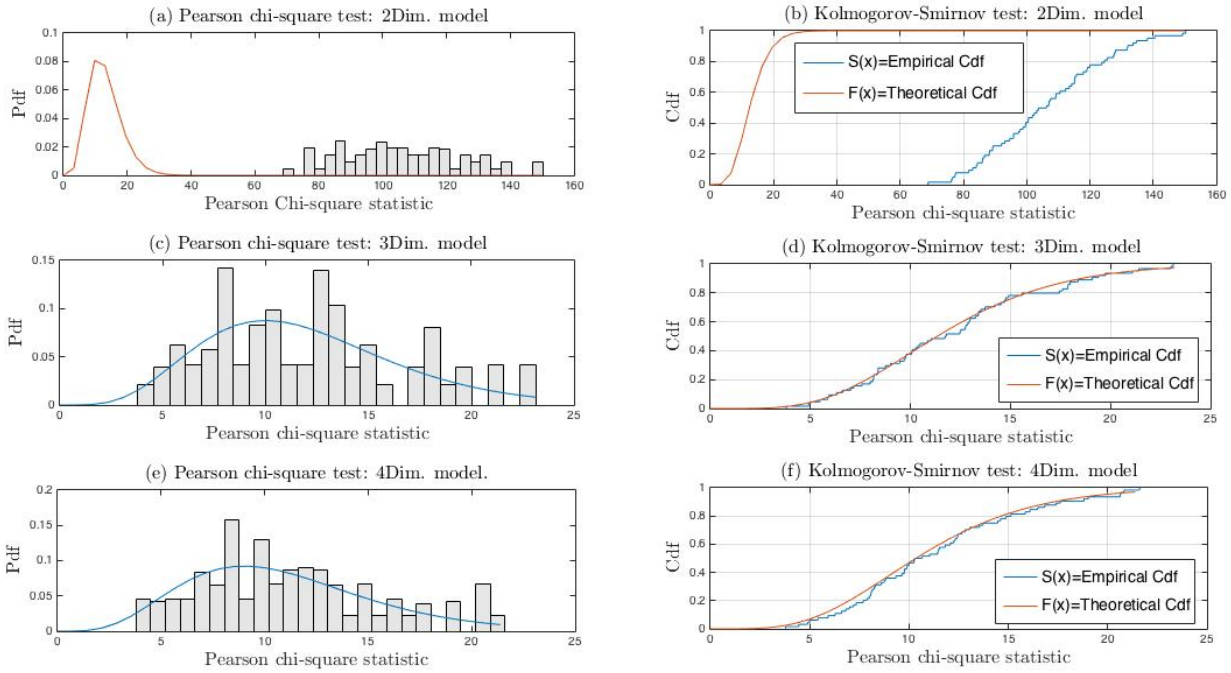

Figure 6.3: Density histograms of the Pearson chi-square and the Kolmogorov-Smirnov goodness of fit tests of the three models on the three dimensional final size epidemic data when $\varepsilon = 0.3$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.0000 T=0.999924	h=0, p= 1.00000, T=0.0000	**h=1**, p=0.0000, T=0.999924
3Dim.	h=0,p=0.107050, T=0.053768	h=0, p=0.120041, T=0.045710	h=0, p=0.053533 T=0.053768
4Dim.	h=0, p=0.102531, T=0.054165	h=0, p=0.059711 T=0.052750	h=0, p=0.051272 T=0.054165

Table 6.6: Summary of the Kolmogorov-Smirnov test for the upper 5% points for the three dimensional final size epidemic data when $\varepsilon = 0.3$ in figures 6.3 (a), (d) and (f).

Same scenario as in figures 6.2 (a)-(f) is observed. The three and four dimensional models are sufficient fit on the three dimensional final size epidemic data when the misclassification probability is large.

From table 6.6, with the misclassification probability $\varepsilon = 0.3$ we see that the null hypothesis is not rejected from the two sided test at 0.05 for the three and four dimensional models owing to insignificant differences between the empirical cumulative distribution functions and the cumulative of the chi-square idistribution function in both directions. They are good approximations.

6.9.3 Table of mean and variance of the Pearson chi-square goodness of fit statistic on the three dimensional final size epidemic data.

Misc. Prob.	2Dim. Model.		3Dim. Model		4Dim. Model	
	Sim. Chi.	Sim. Chi.	Sim. Chi.	Sim. Chi.	Sim. Chi.	Sim. Chi.
	mean	var.	mean	var.	mean	var.
$\varepsilon = 0.0$	12.85	25.875	12.383	24.172	11.562	22.615
$\varepsilon = 0.1$	82.009	288.79	11.853	24.67	11.013	22.714
$\varepsilon = 0.3$	104.9	376.23	11.89	22.162	10.936	20.793

.

Table 6.7: Table of the mean and variance of the Pearson chi-square goodness of fit statistic on the four dimensional final size epidemic data

From table 6.7, we see that the mean and variance of the two dimensional model increases with increasing misclassification probability and therefore not a sufficient fit to the three dimensional final size epidemic data when the misclassification probability is large, which agrees with figures 6.2 (a)-(f) and 6.3 (a)-(f) respectively. Here, 2Dim. is the two dimensional model, 3Dim. is the three dimensional model, 4Dim. is the four dimensional model, Sim. is the simulated mean and variance of the goodness of fit statistics, Misc. Prob. are the misclassification probabilities.

At 5% significance, we see that if $\varepsilon \neq 0$ then the proportion of the simulations rejected for the two dimensional model increases towards 1, as expected in contrast to the behaviour of the three dimensional model in table 7.6. The three and four dimensional models are sufficient on the three dimensional final size epidemic data.

		Proportion Rejected.		
Pearson Chi-square Statistic.	Upper 5% point	$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 0.3$
χ^2_{13}	22.36	0.0480	1	1
χ^2_{12}	21.03	0.0800	0.0640	0.110
χ^2_{11}	19.68	0.0640	0.0960	0.0940

Table 6.8: Table of the proportion of the simulations rejected from the Pearson chi-square test on the final size epidemic data.

6.10 Plots of the mean and variance of the Pearson chi-square goodness of fit statistic on the three dimensional final size data.

In this section, we employed the function and subroutines in section 6.9 to compute and plot the mean and variance of the Pearson chi-square goodness of fit statistic, including the proportion of the simulations rejected at 5% siginficance for a range of $\varepsilon \in [0, 0.1]$ and theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$ respectively. We employed the household structure in [1] but fifty times population size and minimum epidemic size of 1000 to allow the occurrence of large infections in our simulations and hence global epidemic

in the population.

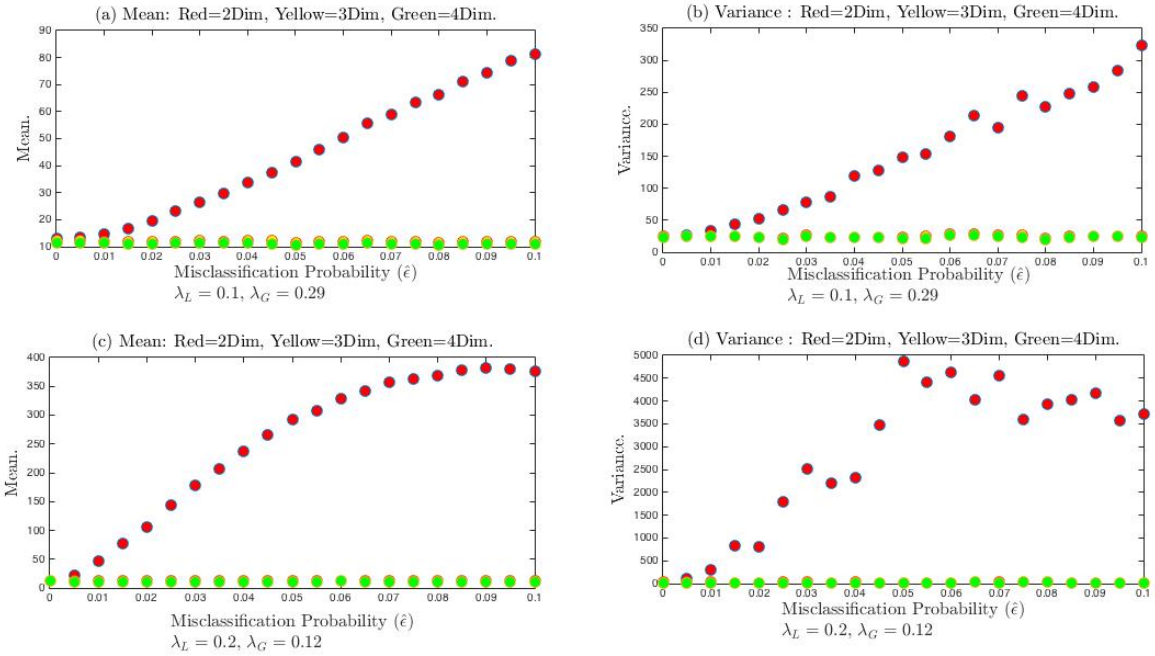

Figure 6.4: Plots of the mean and variance of the Pearson chi-square goodness of fit statistics for the models when $\lambda_L = 0.1$, $\lambda_G = 0.29$ and $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\varepsilon \in [0, 0.1]$, step size of 0.005.

In figures 6.4 (a)-(d), we found that with increasing ε, the mean and variance of the Pearson chi-square goodness of fit statistic from the two dimensional model increases further away from their theoretical counterparts, while those from three and four dimensional models are close to their theoretical counterparts.

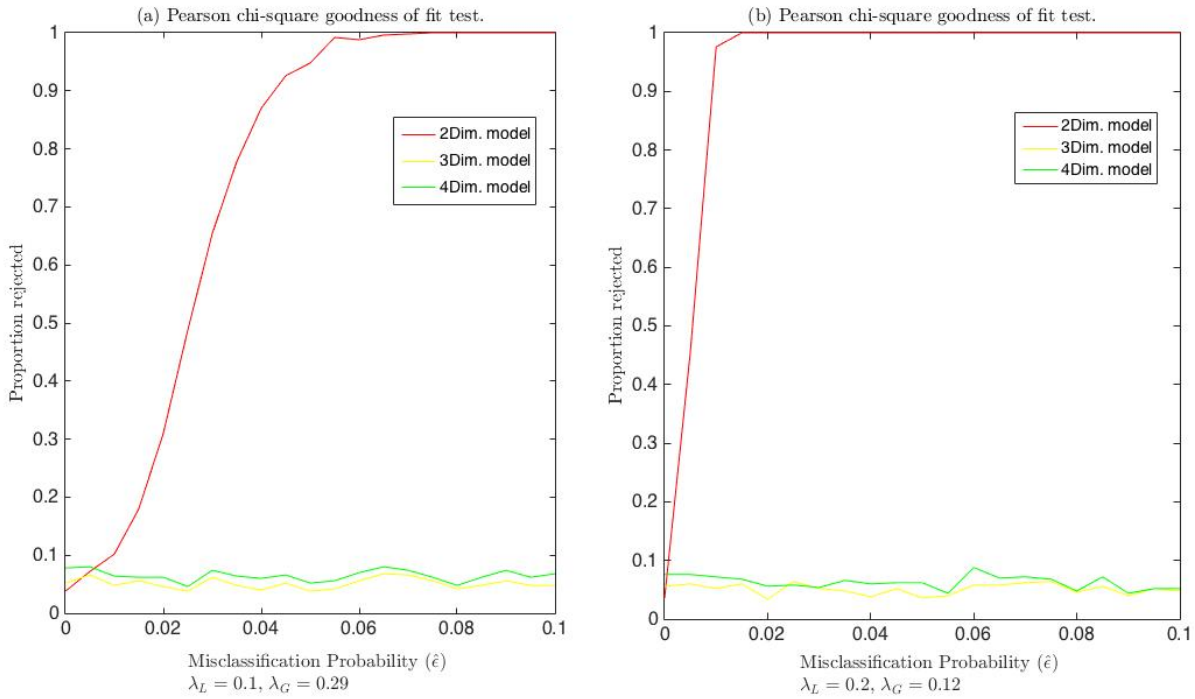

Figure 6.5: Plots of the proportion of the simulations rejected at 5% significance from the Pearson chi-square goodness of fit tests for $\varepsilon \in [0, 0.1]$, step size of 0.005.

In figures 6.5 (a) and (b), we fitted the three models to three dimensional epidemic data using the Pearson chi-square goodness of fit test at the default upper 5% siginficance given by, 22.36, 21.03, and 19.68.

We see that with increasing $\varepsilon \in [0, 0.1]$, the proportion of the simulations rejected from the chi-square goodness of fit test for the two dimensional model increases toward 1 in line with the behaviours of the models in figures 6.4 (a)-(d). The three and four dimensional models have small proportion infected compared to those of the two dimensional model as the misclassification probability increases from zero.

These behaviours agree with our earlier studies, that the three and four dimensional models are the sufficient fit on the three dimensional final size epidemic data.

6.11 The Pearson chi-square goodness of fit tests on the four dimensional final size epidemic data.

We fit the three models to four dimensional final size epidemic data for ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$ in table 6.9 using the following function.

Run the function FourDimThreeATwoSNsimhousesChsqlik to simulate four dimensional final size household epidemic data with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G, vector of household sizes, minimum epidemic size and misclassification probabilities ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$. It then calculates the other corresponding parameters of the three models with Gamma(a, b) infectious period distribution, computes their chi-square statistics, their mean and variance and the porportion of the simulation rejected at 5% significance.

It plots the density histogram of the Pearson chi-square statistics superimposed with their theoretical chi-square distribution function. It also computes and plot the empirical cumulative distribution function of the Pearson chi-square statistic and the cumulative of the hypothesized chi-square distribution function.

These are accomplished with the functions in 5.4.1 and the following subroutines.

a.) final_sizep$(a, b, \pi, n, \lambda_L)$ also computes the Pearson chi-square statistics, associated with the two dimensional model

b.) falseMisclass2a(fneg, fpos, n), computes the Pearson chi-square statistic associated with three dimensional model.

c.) SIRfalsePmisclass$(a, b, n, \pi, \lambda_L,$ fneg, fpos) also computes the Pearson chi-square statistic associated with the four dimensional model

	Serial Number.					
Misclassification Probability.	1	2	3	4	5	6
ε_{FN}	0.0	0.2	0.01	0.02	0.2	0.3
ε_{FP}	0.2	0.0	0.02	0.01	0.3	0.2

Table 6.9: Table of misclassification probabilities 1 to 6.

6.11.1 When the misclassification probabilities are $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$.

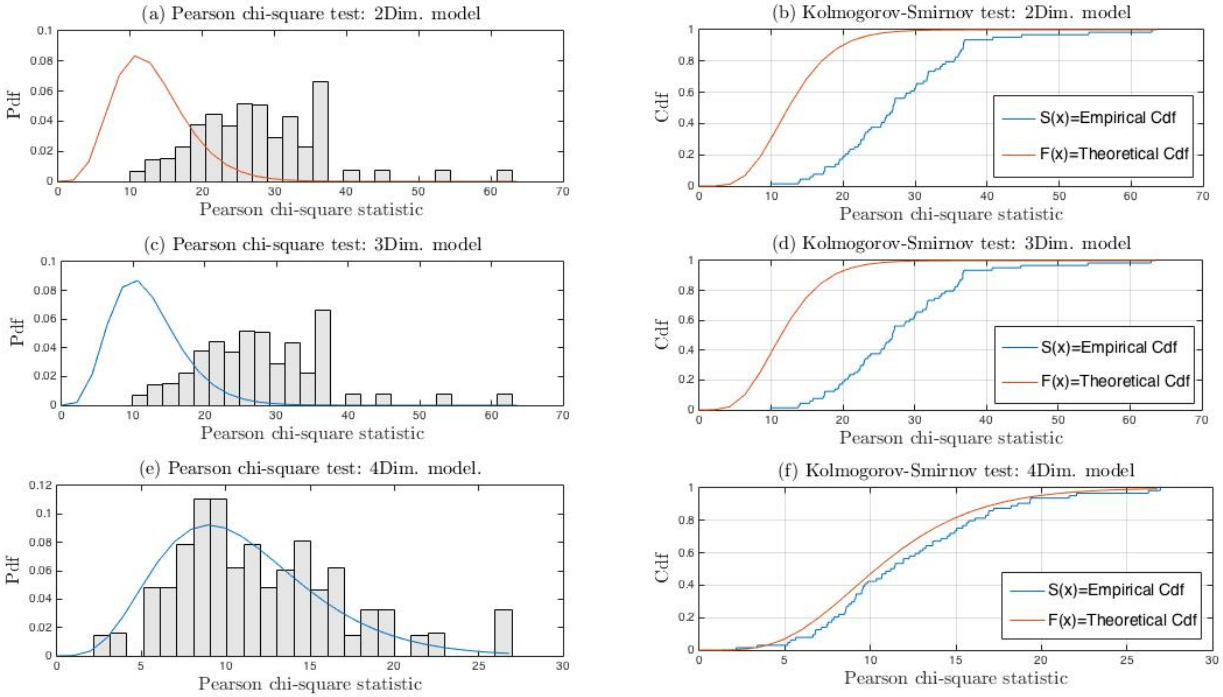

Figure 6.6: Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical chi-square distributions and plots of the empirical cumulative distribution function of the Pearson chi-square goodness of fit statistic with their hypothesized distributions for the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$.

In figures 6.6 (a), (b), (c) and (d), the two and three dimensional models are struggling fitting to the four dimensional final size epidemic data when the misclassification probabilities are far apart from each in line with the discussion in section 5.7. Only the four dimensional model sufficiently fits the four dimensional final size epidemic data under this circumstance as shown by the plots in figures 6.6 (e) and (f).

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.00000, T=0.749446	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=0.749446
3Dim.	**h=1**, p=0.00000, T=0.784782	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000, T=0.78472
4Dim.	**h=1**, p=0.000005, T=0.112961	h=0, p=0.794406, T=0.014844	**h=1**, p=0.000003 T=0.112961

Table 6.10: Summary of the Kolmogorov-Smirnov goodness of fit test with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0.2$ in figures 6.6 (a)-(f).

In table 6.10, we see similar behaviour in table 6.5.

6.11.2 When the misclassification probabilities are $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$.

In figures 6.7 (a) and (c), the two and three dimensional model failed to fit the final size epidemic data. Only the four dimensional model sufficiently fits the four dimensional final size epidemic data given this scenario.

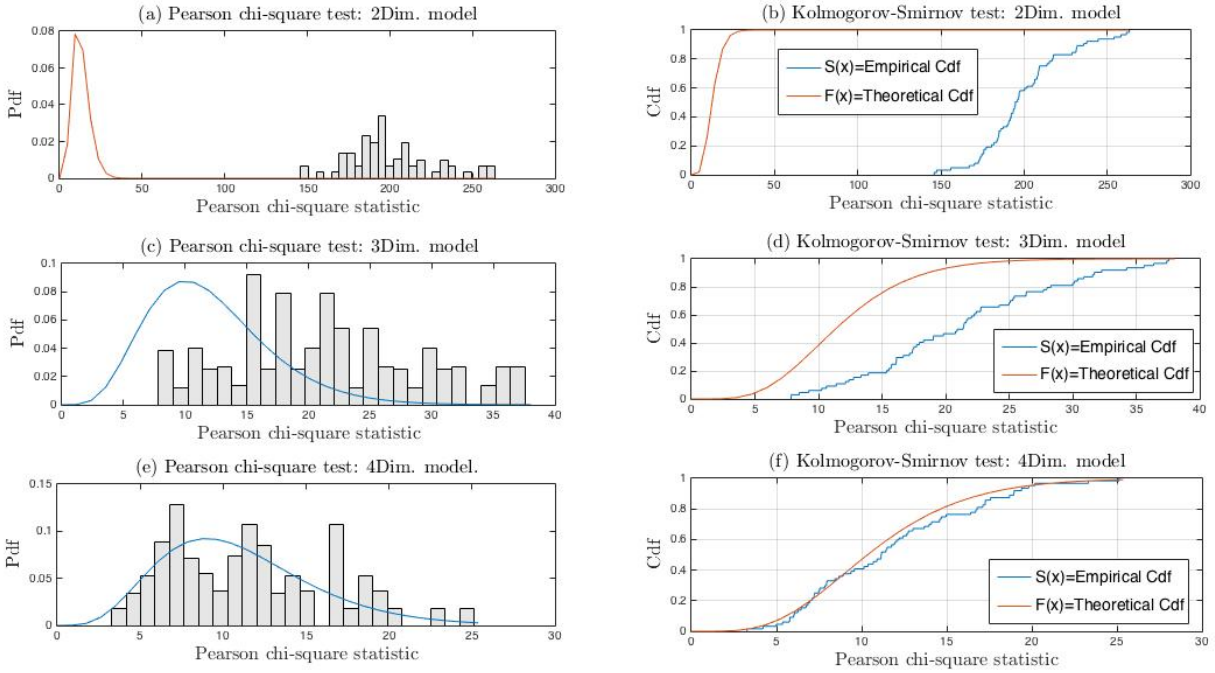

Figure 6.7: Density histogram of the Pearson chi-square, the likelihood ratio chi-squared goodness of fit statistics superimposed with their theoretical chi-square distributions and plots of the empirical cumulative distribution functions with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.00000, T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
3Dim.	**h=1**, p=0.00000, T=0.591303	h=0, p=0.999802 T=0.000181	**h=1**, p=0.00000 T=0.591303
4Dim.	**h=1**, p=0.000011 T=0.109522	h=0, p=0.105348 T=0.047106	**h=1**, p=0.000006 T=0.109522

Table 6.11: Summary of the Kolmogorov-Smirnov goodness of fit tests wit the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0$ in figures 6.7(b), (d) and (f).

In table 6.11, similar behaviour in table 6.5 are observed.

6.11.3 When the misclassification probabilities are $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$

In figures 6.8 (a)-(f), the three model are sufficient fit on the four dimensional final size epidemic data, since the misclassification probabilities are small and close to each other as theoretically expected. Clarity of the models behaviours can be seen from figures 6.8 (b), (d) and (f) with small distances between the empirical cumulative distribution function and their theoretical counterparts. In general, the two, three and four dimensional models are sufficiently fit to the four dimensional final size epidemic data.

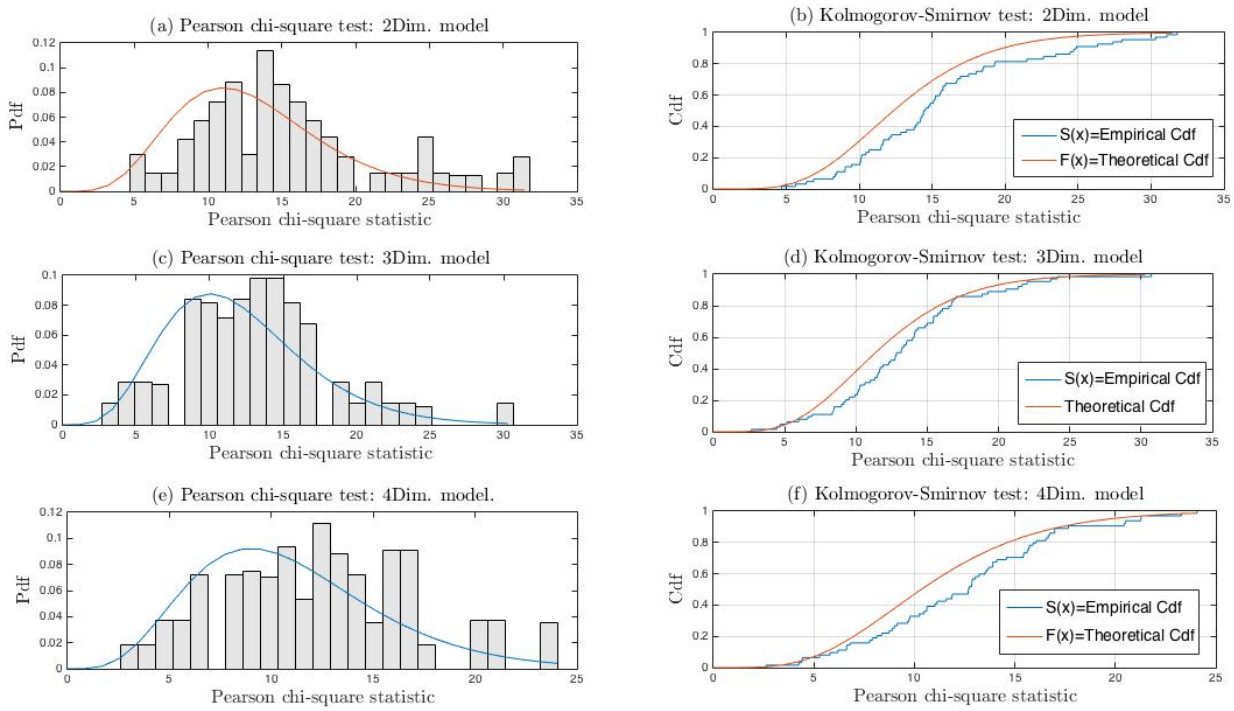

Figure 6.8: Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plot of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.00000 T=0.230967	h=0, p=0.991673 T=0.002579	**h=1**, p=0.00000 T=0.250967
3Dim.	**h=1**, p=0.00000 T=0.169782	h=0, p=0.754431 T=0.016460	**h=1**, p=0.00000 T=0.169782
4Dim.	**h=1**, p=0.00000 T=0.125768	h=0, p=0.059114 T=0.052845	**h=1**, p=0.00000 T=0.125768

Table 6.12: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.01$, $\varepsilon_{FP} = 0.02$ in figures 6.8 (b), (d) and (f).

Similar behaviours in table 6.5 are repeated in Table 6.12.

6.11.4 When the misclassification probabilities are $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$

In figures 6.9 (a)-(f), similar behaviours of the three models in 6.8 (a)-(f) can be seen. However with the false negative misclassification probability larger than the false positive misclassification probability the two dimensional model struggled fitting to the four dimensional final size epidemic data while the three and four dimensional models are sufficient fit to the final size epidemic data. This clarity is shown by the distance between the empirical cumulative distribution function and the cumulative of the hypothesized distribution in figures 6.9 (b), (d) and (f).

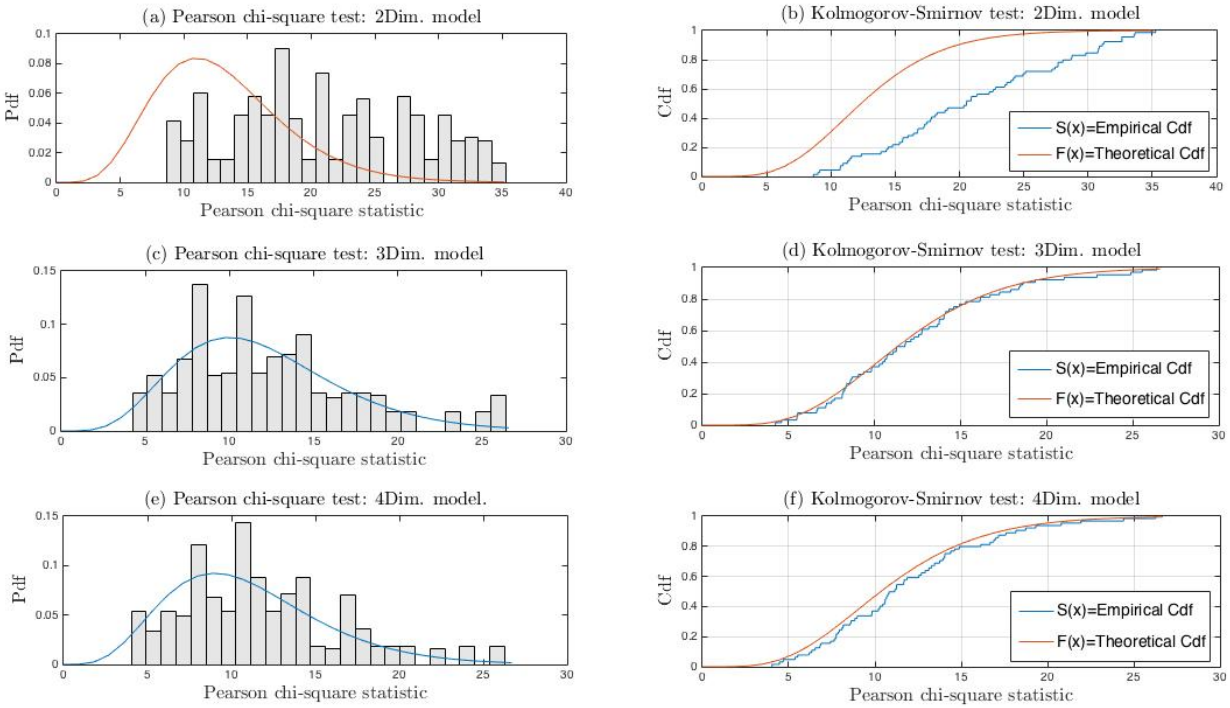

Figure 6.9: Density histograms of the Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, 0.00000 T=0.494793	h=0, p=0.998864 T=0.000772	**h=1**, p=0.00000 T=0.494793
3Dim.	h=0, p=0.089320 T=0.055416	**h=1**, p=0.044664 T=0.055416	h=0, p=0.278133 T=0.035443
4Dim.	**h=1**, p=0.000002 T=0.118042	h=0, p=0.962047 T=0.005897	**h=1**, p=0.000002 T=0.118042

Table 6.13: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$ in figures 6.9 (b), (d) and (f).

In table 6.13 the null hypothesis for the two sided test is rejected from the two and four dimensional models owing to significant discrepancies between their cumulative distribution functions in one direction, while that of the three dimensional model which is small compared to those of the two and four dimensional models is not rejected. This behaviour is associated with the small difference between the misclassification probabilities and makes the four dimensional model approximately equal to the three dimensional model.

6.11.5 When the misclassification probabilities are $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$

In figures 6.10 (c), (d), (e) and (f), in line with our discussion, the three and four dimensional are sufficient fit on the four dimensional final size epidemic data, while the two dimensional model failed to fit the final size epidemic data as in figures 6.10 (a) and (b).

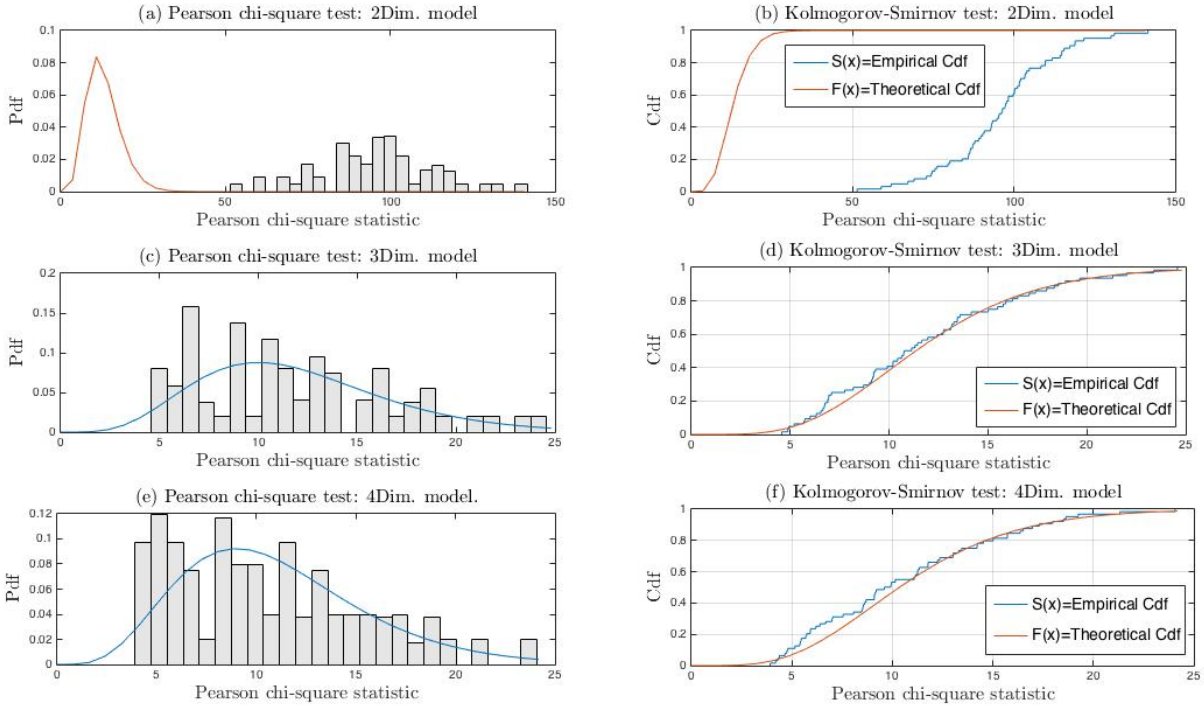

Figure 6.10: Density histograms of Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic data, when $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.00000 T=0.999998	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=0.999998
3Dim.	**h=1**, p=0.000038 T=0.103784	**h=1**, p=0.000019 T=0.103784	h=0, p=0.304117 T=0.034173
4Dim.	**h=1**, p=0.000004 T=0.114439	**h=1**, p=0.000002 T=0.114439	h=0, p=0.306216 T=0.034073

Table 6.14: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.3$ in figures 6.10 (b), (d) and (f).

From table 6.14, the null hypothesis is rejected for the two sided test similar to earlier cases. The empirical cumulative distribution functions from the three and four dimensional models are good approximations.

6.11.6 When the misclassification probabilities are $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$

In figures 6.11 (a)-(f), we see similar behaviour in figures 6.10 (a)-(f), in which the two dimensional model struggled fitting the four dimensional final size epidemic data when the misclassification probabilities are not close to 0 as expected.

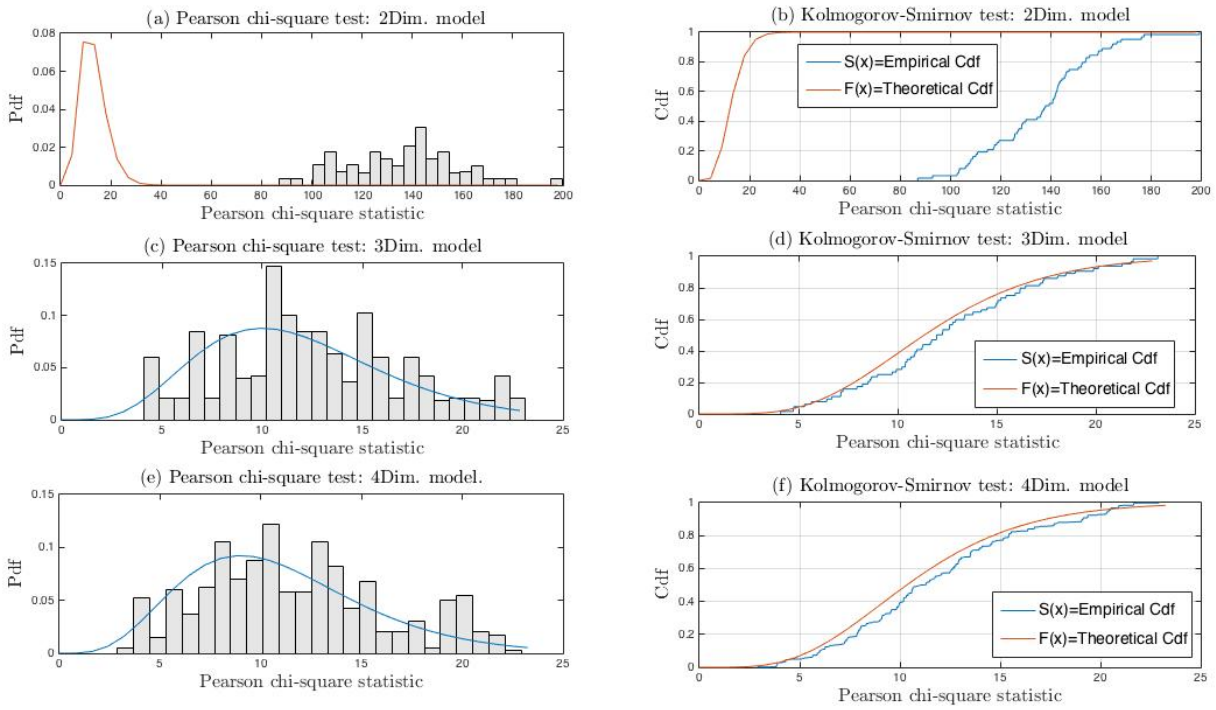

Figure 6.11: Density histograms of Pearson chi-square goodness of fit statistics superimposed with their theoretical counterparts and plots of the empirical cumulative distribution functions of the Pearson chi-square goodness of fit statistics with their theoretical counterparts of the three models on the four dimensional final size epidemic, when $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$.

Model	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
2Dim.	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
3Dim.	**h=1**, p=0.00000 T=0.109797	h=0, p=0.563241 T=0.023632	**h=1**, p=0.000005 T=0.109797
4Dim.	h=0, p=0.072178 T=0.057294	h=0, p=0.381282 T=0.030724	**h=1**, p=0.036091 T=0.057294

Table 6.15: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points for the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$ in figures 6.11 (b), (d) and (f).

In table 6.15, the null hypothesis of the two sided test from the four dimensional model is not rejected at 0.05 significance because of the small discrepancy between the cumulative distribution functions in one direction. The empirical cumulative distribution function from the four dimensional model is a better approximation of the cumulative of the chi-square distribution function.

6.12 Table of mean and variance of the Pearson chi-square goodness of fit statistics of the three models on the four dimensional final size epidemic data.

In table 6.16, we presented the mean and variance of the Pearson chi-square goodness of fit statistic for the three models. We see that the four dimensional model is the best fit to four dimensional final size epidemic data especially when the misclassification probabilities are significantly large and far apart from each other.

Here, 2Dim=two dimensional model, 3Dim=three dimensional model, 4Dim=four dimensional model, Misc. Prob.=misclassification probabilities, Sim. chi. mean=simulated mean of the Pearson chi-square statistic, while Sim. chi. var is the corresponding variance. These values are compared with their theoretical counterparts in table 6.2.

Misc. Prob.	2Dim. Model.		3Dim. model.		4Dim. model.	
	Sim. chi. mean	Sim. chi. var	Sim. chi. mean	Sim. chi. var	Sim. chi mean	Sim. chi. var
$\varepsilon_{FN} = 0.0$, $\varepsilon_{FP} = 0.2$	26.82	85.458	26.82	85.458	11.714	25.028
$\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.0$	2004.54	802.49	20.413	54.791	11.695	23.091
$\varepsilon_{FN} = 0.01$, $\varepsilon_{FP} = 0.02$	14.331	33.625	12.046	25.41	11.315	24.251
$\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$	20.641	61.088	11.808	25.394	11.16	24.055
$\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.3$	95.526	324.01	12.401	24.165	11.113	21.21
$\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$	140.3	455.53	12.763	22.885	11.359	20.475

Table 6.16: Table of mean and variance of the Pearson chi-square goodness of fit statistics on the four dimensional final size epidemic data.

Pear. Chi. Stat.	Upper 5% point	Proportion Rejected.			
		$\varepsilon_{FN} = 0.$ $\varepsilon_{FP} = 0.2$	$\varepsilon_{FN} = 0.2$ $\varepsilon_{FP} = 0$	$\varepsilon_{FN} = 0.2$ $\varepsilon_{FP} = 0.3$	$\varepsilon_{FN} = 0.3$ $\varepsilon_{FP} = 0.2$
χ^2_{13}	22.36	0.604	1	1	1
χ^2_{12}	21.03	0.698	0.502	0.064	0.042
χ^2_{11}	19.68	0.060	0.064	0.032	0.078

Table 6.17: Table of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test for misclassification probabilities in $[0, 0.5)$.

Table 6.17 provides further insight into the misfit of the two dimensional model when the misclassification probabilities are not close to 0. In such situations, the proportion of the simulations rejected is exactly 1 and theoretically signifies the model misfit to the four dimensional final size epidemic data. While the three dimensional model sufficiently fits the four dimensional final size epidemic if the misclassification probabilities are close to each other, otherwise it struggles fitting to the four dimensional final size epidemic data as in table 6.17 for $\varepsilon_{FN} = 0, \varepsilon_{FP} = 0.2$ and vice versa and also has high proportion of the simulations rejected compared to those of the four dimensional model.

6.13 Plots of the mean and variance of the Pearson chi-square goodness of fit statistic.

Using simulation studies we explored the parameter estimates along the diagonals of the misclassification probabilities region, $[0, 0.5)$, with the line $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$, step size of 0.01 and compute the Pearson chi-square goodness of fit statistics of three model, their mean, variance and the proportion of the simulation rejected at 5% significance.

We present results of the studies for ε_{FN}, $\varepsilon_{FP} \in [0, 0.2]$, as the behaviour of the mean and variance of the Pearson chi-square goodness of fit statistic are repeated over the remaining part of the permissible region, ε_{FN}, $\varepsilon_{FP}(0.2, 0.5]$ for theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$ respectively.

These computations are implemented using the following function and subroutines.

Run the function, FourThreeTwoDonFourfposChsqlik to simulate four dimensional household epidemic with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G, misclassification probabilities $\varepsilon_{FP} \in [0, 0.5)$. It explores the parameter estimates of the three models along the line $\varepsilon_{FN} = \alpha - \varepsilon_{FP}$ of the misclassification probabilities region, where α is defined as $\varepsilon_{FP} \in [0, \alpha]$, $\alpha < 0.5$. It then plot the mean and variance of the Pearson chi-square statistics and those of the chi-square difference statistic for the three models.

It also computes and plot the proportion of the simulations rejected from the Pearson chi-square and the chi-square difference statistics at 5% significance. These are accomplished using the subroutines in section 6.7.

6.13.1 Exploring the estimates along the diagonals, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$, theoretical parameters corresponding to $z = 0.7298, 0.2144$ respectively.

We implement the function and subroutines in section 6.13 with minimum epidemic threshold of 1000 and household structure $[133, 189, 108, 106, 31] * 50$ in figures 6.12 (a)-(d) and figures 6.13 (a) and (b) respectively.

In figures 6.12 (a)-(d), the mean and variance of the Pearson chi-square goodness of fit

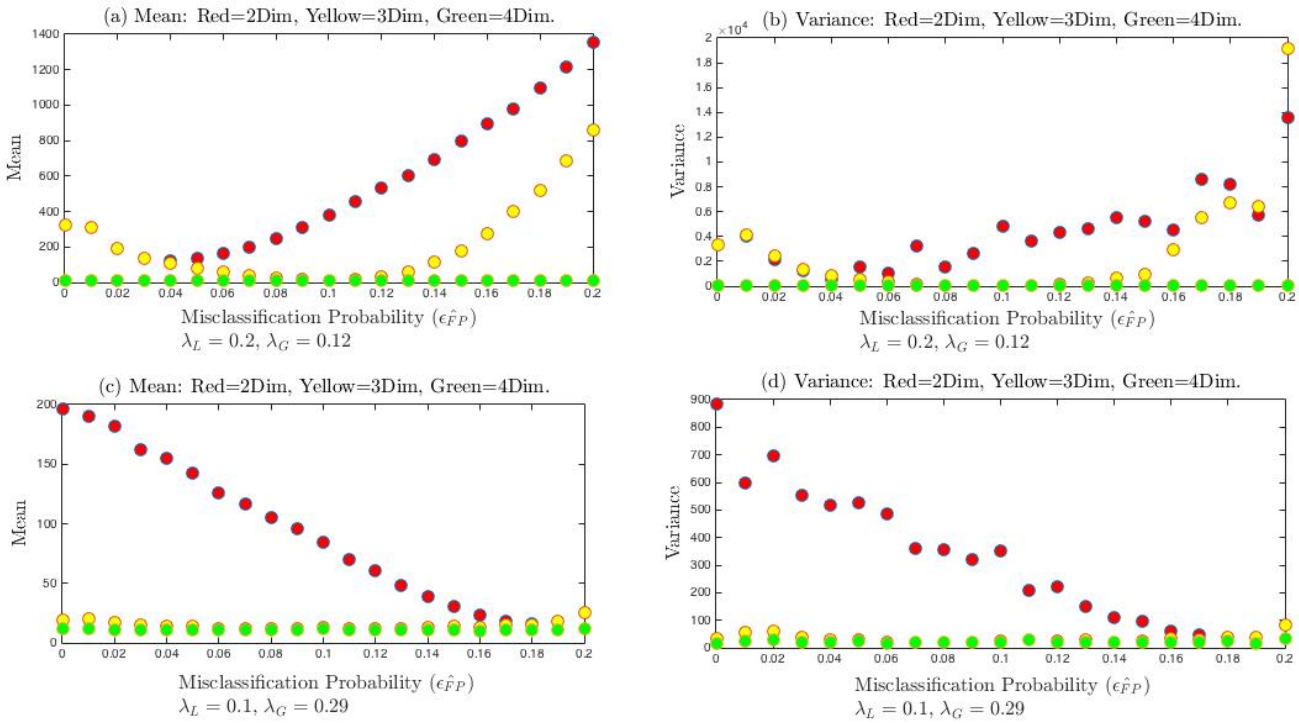

Figure 6.12: Plots of the mean and variance of the chi-square goodness of fit statistics of the models, when the estimates are explored along the diagonals of the misclassification region $[0, 0.2]$, with step size of 0.01 for theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$ respectively.

statistics for the three models are computed by exploring the estimates along the diagonal of the misclassification probabilities, $\varepsilon_{FP} \in [0, 0.2]$ for $z = 0.2144, 0.7298$.

We see that with theoretical parameters corresponding to $z = 0.7298$, the mean and variance of the Pearson chi-square goodness of fit statistics of the three and four dimensional models are consistent and are approximately equal to their theoretical counterparts, while those of the two dimensional model tends toward their theoretical mean and variance. Also, inconsistent behaviour of the mean and variance of the two and three dimensional models can be seen for $\varepsilon_{FP} \in [0, 0.2]$ when $\lambda_L = 0.2$ and $\lambda_G = 0.12$.

When $\varepsilon_{FP} = 0.1$, the false negative probability, $\varepsilon_{FN} = 0.1$ and hence the four dimensional model reduces to the three dimensional model with the mean and variance close to the theoretical counterpart for the two set of theoretical parameters.

Thus no matter the choice of theoretical parameters corresponding to z, the mean and variance of the four dimensional model are consistently stable and close to their theoretical counterpart, while those of the three dimensional model are stable for $\varepsilon_{FP} = 0.1$ and theoretical parameters corresponding to large values of $z \in [0, 1]$. Those of the two dimensional model are not reliable.

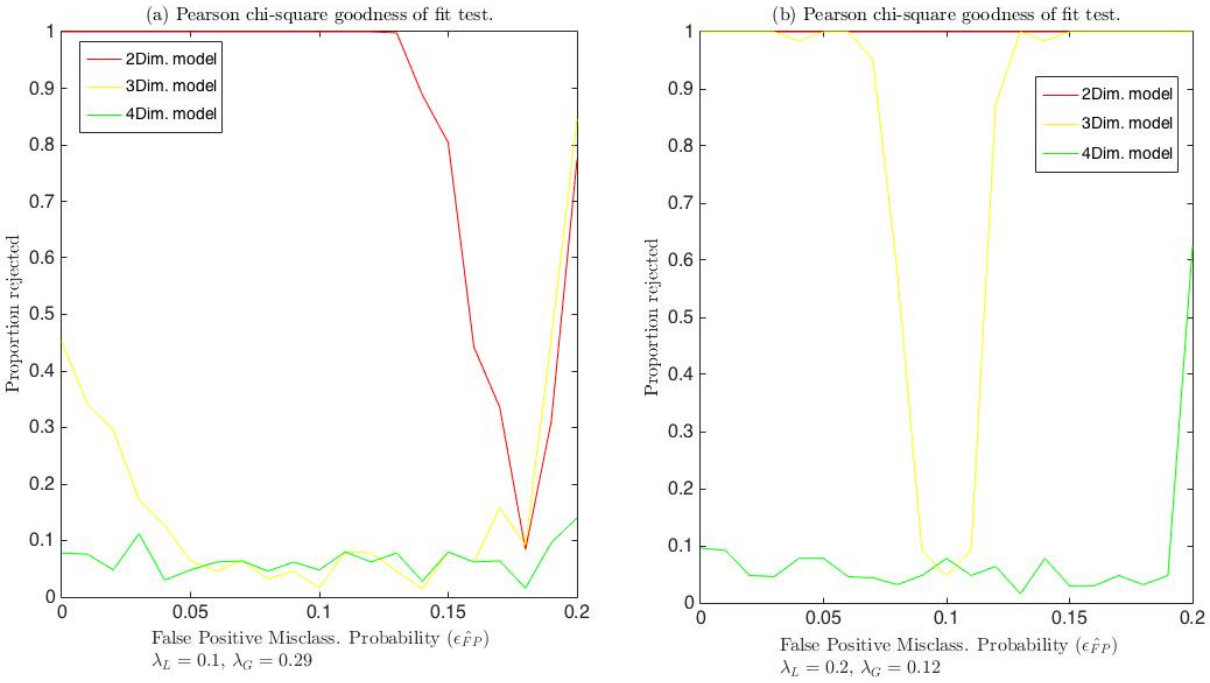

Figure 6.13: Plots of the proportion of the simulations rejected from the Pearson chi-square goodness of fit test to our dimensional final size epidemic data for theoretical parameters corresponding to $z = 0.2144$. and $z = 0.7298$ respectively.

Further clarity on the behaviours of the models in 6.12 are provided in figures 6.13 (a) and (b). For the two dimensional model, the proportion of the simulations rejected are influenced by the magnitude of the theoretical parameters corresponding to z.

For example when the theoretical parameters corresponds to $z = 0.2144$, the proportion of the simulation rejected is consistently 1 so also are those of the three dimensional model except when $\varepsilon_{FP} = 0.1$ with the proportion rejected approximately equal to 0.05 as expected. Those of the four dimensional model are stable and close to the required proportion rejected

at 5% significance.

Thus when the theoretical parameters corresponds to $z = 0.2144$, the three dimensional model is sufficient on the four dimensional final size epidemic data, while the two dimensional model is not.

Also when $\varepsilon_{FP} \in [0, 0.2]$ and theoretical parameters corresponds to $z = 0.7298$, the proportion of the simulations rejected from the Pearson chi-square goodness of fit test for the two dimensional model is approximately 1, while those from the three and four dimensional models are less than 1 as theoretically expected.

6.13.2 Exploring the estimates along the vertical axis of the misclassification probability region.

We implement the procedures with the function, FourThreeTwoDonFourNonGraphSNsimhousesSIRchiqlik2 and subroutines as follows.

Run the function, FourThreeTwoDonFourNonGraphSNsimhousesSIRchiqlik2 to simulate four dimensional household epidemic data with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G and misclassification probabilities ε_{FN}, $\varepsilon_{FP} \in [0, 0.5)$.

Here one of the misclassification probability is held fixed, while the other is varied in $[0, 0.5)$.

The function then estimate the parameters, computes and plot the root mean square error. It also computes the Pearson chi-square statistic and the chi-square difference statistic for the three models and also plot their mean and variance and the proportion of the simulation rejected from the Pearson chi-square goodness of fit test and those of the the chi-square difference tests at 5% significance using the subroutines in sections 6.7

We implement these procedures with theoretical parameters, corresponding to $z = 0.7298$, 0.2144 and then other corresponding parameters are estimated along the vertical, where $\varepsilon_{FP} = 0.01$ and ε_{FN} in $[0, 0.2]$ with step size of 0.01 as in figures 6.14 (a)-(d).

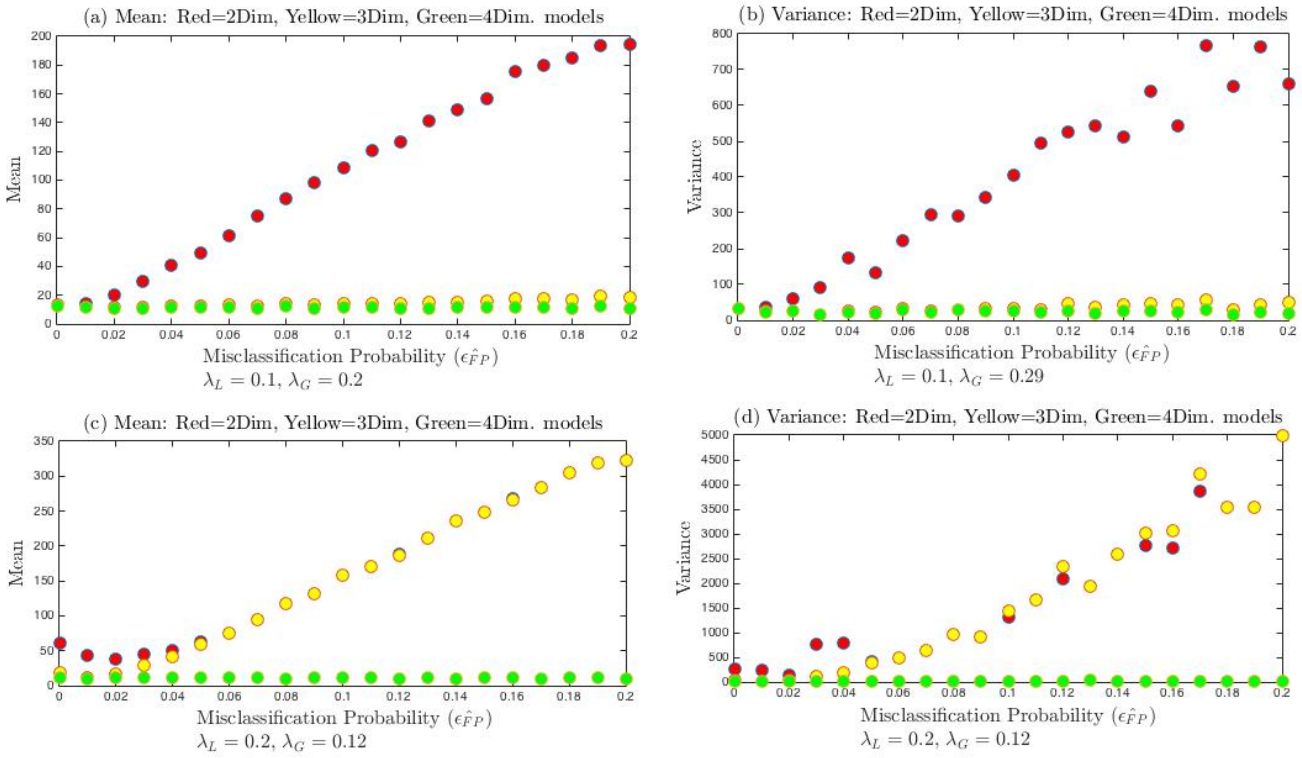

Figure 6.14: Plots of the mean and variance of the chi-square goodness of fit statistics for the three models with $\varepsilon_{FP} = 0.01$ while varying $\varepsilon_{FN} \in [0, 0.2]$ with step size of 0.01 and theoretical parameters corresponding to $z = 0.7298$ and z=0.2144.

In Figures 6.14 (a)-(d), we explored the estimates of the model parameters along the vertical axis of the misclassification probabilities with $\varepsilon_{FP} = 0.01$ and $\varepsilon_{FN} \in [0, 0.2]$ for theoretical parameters corresponding to $z = 0.2144$, 0.7298 as follow.

Firstly, with theoretical parameters corresponding to $z = 0.7298$, the mean and variance of the Pearson chi-square statistics of the three and four dimensional models are consistently stable and approximately equal to their theoretical counterparts as the misclassification probabilities are varied in $[0, 0.2]$, while those of the two dimensional model are unstable.

For example the mean and variance of the Pearson chi-square goodness of fit statistics of the two and three dimensional models with theoretical parameters corresponding to $z = 0.2144$, are approximately equal to each other and unstable for $\varepsilon_{FN} \in [0, 0.2]$, while those of the four dimensional model remains consistently stable and approximate its theoretical

counterpart.

Also, those of the three and four dimensional models with theoretical parameters corresponding $z = 0.7298$, are stable and are approximately equal to their theoretical counterparts, while those of the two dimensional model are unstable with increasing $\varepsilon_{FN} \in [0, 0.2]$.

In line with our studies in section 5.4.1, 5.6.1 and 5.7 and those in figures 6.12 (a)-(d), we see that no matter the choice of the theoretical parameters with corresponding $z \in [0, 1]$, the estimates of the four dimensional model are more precise than those of the two and three dimensional models when the misclassification probabilities are far apart from each other and therefore outperforms them on the four dimensional epidemic data.

6.13.3 Exploring the estimates along the horizontal axis of the misclassification probability region.

Using the same theoretical parameters in subsection 6.13.2, we simulated four dimensional epidemic data and explored the estimates along the horizontal axis of the misclassification proabbilities with $[\varepsilon_{FP} \in 0, 0.2]$. Here we fixed $\varepsilon_{FN} = 0.01$ and vary $\varepsilon_{FP} \in [0, 0.2]$.

The Pearson chi-square goodness of fit, their mean and variance are then computed and ploltted for the three three models in figures 6.15 (a)-(d).

Similar behaviours to figures 6.14 (a)-(d) are observed.

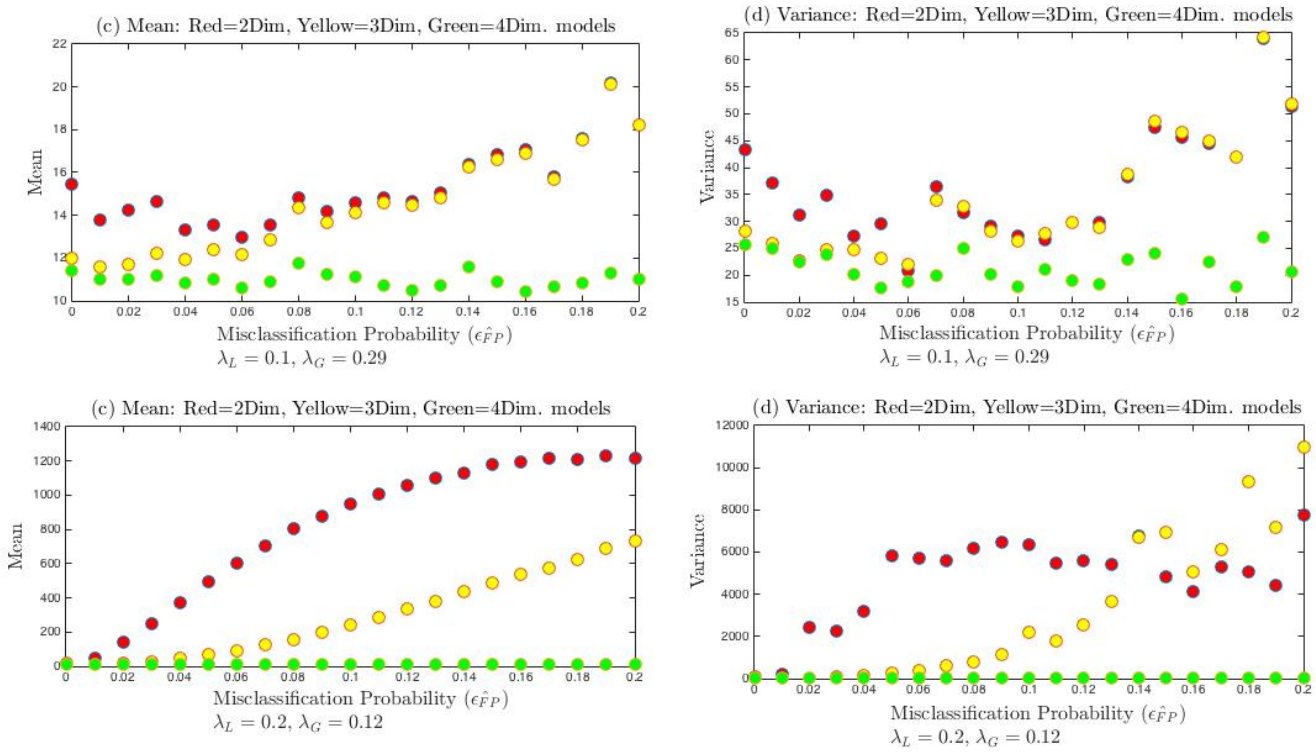

Figure 6.15: Plots of the mean and variance of the chi-square goodness of fit statistics for the three models with $\varepsilon_{FN} = 0.01$ while varying $\varepsilon_{FP} \in [0, 0.2]$ with step size of 0.01 and theoretical parameters corresponding to $z = 0.7298$ and z=0.2144.

6.14 Fitting the three models to [1] Tecumseh Michigan Influenza A(H3N2) epidemic data.

We analysed [1] Tecumseh Michigan final epidemic data with Gamma$(2, 2.05)$ infectious period distribution using the Pearson chi-square goodness of fit test and the function, Addy as follows.

Run the function Addy to estimate the parameters of the three models and compute the Pearson chi-square statistic using subroutines in section 6.7.

These are implemented with [1] household epidemic data in table 6.18.

Estim., Stat. and P-values.	2Dim. Model	3Dim. model	4Dim. model
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.044638	0.044638	0.044638
$\hat{\pi}$	0.86738	0.86738	0.86738
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0$
Pearson Ch-sq Stat.	14.435	14.435	14.435
P-values	P $= 0.3439$	P $= 0.2738$	P $= 0.2099$

Table 6.18: Table of the parameter estimates from [1] final size epidemic data for the three models, the Pearson chi-square goodness of fit statistic and the corresponding P-values for the tests.

6.15 Analyses of the Seattle influenza datasets.

The observed distributions of the 1975-1976 B(H1N1) and 1978-1979 A(H1N1) Seattle influenza epidemic in [28] also discussed in [49] given in tables 6.19 and 6.20 respectively are analysed in two ways. Namely by assuming no misclassification error in the data and hence considering them as two dimensional final size data. The two dimensional model is then fitted to the two datasets by assuming Gamma$(k, 4.1/k), k = 1, 2, 5$ infectious period distributions and analysed in tables 6.22 and 6.21 respectively.

Secondly, fitting the three models to the epidemic datasets for Gamma$(k, 4.1/k), k = 1, 2, 5$ infectious period distributions and compute their Pearson chi-square goodness of fit statistics.

In this way the misclassification probabilities are estimated if they are nonzero and hence provides clarity about the true dimension of the datasets and the model that fits significantly better to the final size epidemic datasets.

Household Size	Number Infected in Household					
	0	1	2	3	4	5
1	9	1	-	-	-	-
2	12	6	2	-	-	-
3	18	6	3	1	-	-
4	9	3	4	3	0	-
5	4	3	0	2	0	0

Table 6.19: Influenza B(H1N1) 1975-1976 final size data.

Household Size	Number infected in household			
	0	1	2	3
1	15	11	-	-
2	12	17	21	-
3	4	4	4	5

Table 6.20: Influenza A(H1N1) 1978-1979 final size data.

6.15.1 Analyses of the epidemic datasets.

If we assume no misclassification of the final size data, then for Gamma$(k, 4.1/k), k = 1, 2, 5$ infectious period distributions, the estimates from the two dimensional models are obtained in tables 6.19 and 6.20,

Parameters	Gamma infectious period distribution		
	Gamma$(1, 4.1)$	Gamma$(2, 4.1/2)$	Gamma$(5, 4.1/5)$
$\hat{\lambda_L}$	0.035083	0.035228	0.035216
$\hat{\lambda_G}$	0.207147	0.204628	0.2031089
$\hat{\pi}$	0.83305	0.83449	0.83536
$\hat{z}$	0.215073	0.215662	0.216028
$\hat{R_*}$	1.1591	1.1613	1.1628

Table 6.21: Estimates from the 1975-1976 Seattle B(H1N1) influenza epidemic.

Parameters	Gamma infectious period distribution		
	Gamma$(1, 4.1)$	Gamma$(2, 4.1/2)$	Gamma$(5, 4.1/5)$
$\hat{\lambda_L}$	0.10876	0.098711	0.092883
$\hat{\lambda_G}$	0.274861	0.274462	0.274388
$\hat{\pi}$	0.53779	0.53828	0.5384
$\hat{z}$	0.550417	0.550412	0.550362
$\hat{R_*}$	1.5562	1.5582	1.5594

Table 6.22: Estimates from the 1978-1979 Seattle A(H1N1) influenza epidemic.

6.16 Fitting the three models to the Seattle household epidemic data.

Having analysed the epidemic datasets with Gamma$(k, 4.1/k)$, $k = 1, 2, 5$ infectious period distributions and assuming no misclassification error in the datasets, which may not be the case, we then explored them for misclassification errors in order to get the appropriate model fit on each of the epidemic dataset. This is achieved using the Pearson chi-square goodness of fit statistics in subsections 6.16.1 and 6.16.2 respectively.

6.16.1 The 1975-1976 Seattle B(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.035083	0.035083	0.0992
$\hat{\pi}$	0.83305	0.83305	0.7827
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0.3828$, $\varepsilon_{FP} = 0.000$
Pear. ch-sq. statistic and P-value	$X_2 = 6.7457$, P=0.9148	$X_3 = 6.7457$, P=0.8740	$X_4 = 4.8383$, P=0.9387

Table 6.23: Parameter estimates and Pearson chi-suare goodness of fit statistics with Gamma$(1, 4.1)$ infectious period distribution from the 1975-1976 B(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.035228	0.035228	0.0750
$\hat{\pi}$	0.83449	0.83449	0.7911
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0.3324$, $\varepsilon_{FP} = 0.0001$
Pear. ch-sq. Statistic and P-value	$X_2 = 6.1456$, P=0.9407	$X_3 = 6.1456$, P= 0.9086	$X_4 = 4.89758$, P= 0.9360

Table 6.24: Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(2, 4.1/2)$ infectious period distribution from the 1975-1976 B(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.035216	0.035216	0.0629
$\hat{\pi}$	0.83536	0.83536	0.7982
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0.2875$, $\varepsilon_{FP} = 0$
Pear. ch-sq. statistic and P-value	$X_2 = 5.797$, P=0.9532	$X_3 = 5.797$, P=0.9260	$X_4 = 4.932282$, P=0.9344

Table 6.25: Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(5, 4.1/5)$ infectious period distribution from the 1975-1976 B(H1N1) influenza epidemic.

6.16.2 The 1978-1979 Seattle A(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.10876	0.10876	0.1088
$\hat{\pi}$	0.53779	0.53779	0.5378
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0$
Pear. ch-sq. statistic and P-value	X_2 =2.0409, P=0.7282	X_3=2.0409, P=0.5640	X_4=2.0409, P=0.3604

Table 6.26: Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(1, 4.1)$ infectious period distribution from the 1978-1979 A(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.098711	0.098711	0.1023
$\hat{\pi}$	0.53828	0.53828	0.5523
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0.0258$
Pear. ch-sq. statistic and P-value	X_2=2.0988, P= 0.7176	X_3=2.0988, P=0.5522	X_4=2.0978, P= 0.3503

Table 6.27: Parameter estimates andPearson chi-square goodness of fit statistics with Gamma$(2, 4.1/2)$ infectious period distribution from the 1978-1979 A(H1N1) influenza epidemic.

Estim., Stat. and P-values.	2Dim. Model.	3Dim. model.	4Dim. model.
	Parameter Estimate	Parameter Estimate	Parameter Estimate
$\hat{\lambda_L}$,	0.092883	0.092883	0.1031
$\hat{\pi}$	0.5384	0.5384	0.5794
Misclass. Prob. Estim.	0	$\varepsilon_{FN} = \varepsilon_{FP} = \varepsilon = 0$	$\varepsilon_{FN} = 0\, \varepsilon_{FP} = 0.0717$
Pear. ch-sq. statistic and P-value	X_2=2.1629, P= 0.7058	X_3=2.1629, P= 0.5393	X_4=2.15002, P= 0.3413

Table 6.28: Parameter estimates and Pearson chi-square goodness of fit statistics with Gamma$(5, 4.1/5)$ infectious period distribution from the 1978-1979 A(H1N1) influenza epidemic.

6.17 Discussion and Comments.

Once there is no misclassification error in the final size epidemic data, then the best model fit to the two dimensional final size data is the two dimensional model. This property is demonstrated in figures 6.1 and table 6.2. The Pearson chi-square and the likelihood chi-squared goodness of fit statistics from the models, are well fitted to their theoretical chi-square distributions, their mean and variance are approximately close to those of its theoretical counterparts. Therefore, a model with smaller number of parameters is preferred, making the two dimensional model the appropriate model fit to two dimensional final size epidemic data if $\varepsilon = 0$.

However, if ε is far from 0 then the two dimensional model begins to struggle fitting to three dimensional final size data as shown in figure 6.2 and 6.3 when $\varepsilon = 0.1$ and $\varepsilon = 0.3$ respectively.

Hence, with increasing ε it becomes unreliable to use the two dimensional model aas the

estimates are known to be biased. The Pearson chi-square statistics have disproportionate mean and variance and hence does not fit its theoretical counterpart as shown in figures 6.2, 6.3 and table 6.7. The three and four dimensional models still provide good fit to the theoretical chi-square distribution in the face of increasing values of the misclassification probabilities as seen in figures 6.2 and 6.3.

However, with large and different misclassification probabilities far apart from each other, the four dimensional model have precise estimates and therefore outperforms the two and three dimensional models on the four dimensional final size epidemic data as demonstrated in tables, 6.2, 6.7 and 6.16 respectively, showing the appproximate mean and variance of the chi-square goodness of fit statistic to their theoretical counterparts of three models.

In general, we have seen that given any choice of $z \in [0, 1]$, the mean and variance of the Pearson chi-square goodness of fit statistics of the four dimensional model are approximately equal to their theoretical mean and variance.

Thus, with increasing misclassification probabilities, the two and three dimensional models will begin to struggle fitting to the four dimensional final size data, with disproportionate parameter estimates and hence poorly fitted density histograms of the Pearson chi-square and likelihood ratio chi-squared goodness of fit statistics of the two models to its theoretical counterparts. These behaviours are exhibited in section 6.11.

Also in table 6.18, the Pearson chi-square goodness of fit statistic of the three models on the [1] final size epidemic data are the same, so also are those of the likelihood ratio chi-squared statistics, with corresponding P-values, given their degrees of freedom 13, 12, and 11, as $\mathrm{P} = \mathrm{P}(\chi^2 \geq \chi^2_{13}) > 0.25$, $\mathrm{P} = \mathrm{P}(\chi^2 \geq \chi^2_{12}) > 0.25$ and $\mathrm{P} = \mathrm{P}(\chi^2 \geq \chi^2_{11}) > 0.100$ respectively, which are the same as those of the likelihood ratio Chi-squared statistic test with the same degrees of freedom.

The observed chi-square goodness of fit statistic are smaller than the critical values at P equal to the P-values for the given degrees of freedom and hence the tests are insignificant. The models fitted are sufficient to the final size epidemic data. In conclusion, the two dimensional model, which is the simplest of the three models is the appropriate model fit to the two dimensional final size epidemic data from the Tecumseh Michigan influenza A(H3 N2)

epidemic data.

Also, from the analyses of the Seattle $1975-1976$ B(H1N1) and $1978-1979$ A(H1N1) influenza epidemics in subsections 6.16.1 and 6.16.2, we see that the misclassification probabilities are estimated as 0 by the three dimensional model contrary to their nonzero estimates given by the four dimensional model. However from their Pearson chi-square goodness of fit statistics, we see that at 5% significance the three models fit sufficiently to the epidemic data.

Chapter 7

Hypothesis test between the models.

7.1 Introduction.

The three models discussed so far are nested within each other, such that each simpler model is obtained by fixing or eliminating parameter in the more complex models. That is, fixing the misclassification probabilities in the four dimensional model to be equal, leads to the three dimensional model, while fixing the misclassification probability in the three dimensional model to be zero leads to the two dimensional model.

The three models can then be compared with regards to their fitness to the final size epidemic data, using chi-square difference test, in which the difference between the Pearson chi-square goodness of fit statistic from the models are evaluated and analysed. If the difference is significant then the model with more estimated parameters fits the final size epidemic data better than the smaller model with less parameters.

These procedures provide information as to the need to estimate the additional parameter and to employ the model with larger number of parameters. However, if the chi-square difference statistic is insignificant, then the more complicated model does not offer significant improvement over the one with smaller number of parameters. The parameter in question is then ignored by putting it equal to zero. In any case, this procedure allows us to decide whether a given model fits significantly better than the other competing models.

Sometimes, there may be a need to employ more than one goodness of fit test in order

to provide clarity on the fitness of the models to the final size epidemic data. We have also employed the Kolmogorov-Smirnov goodness of fit test discussed in section 6.5 to look at the quality of the Pearson chi-square approximations.

7.2 Chi-square difference test.

Using simulation studies, we fitted the three models to final size epidemic data and computed their Pearson chi-square goodness of fit statistic, X_2, X_3, X_4 respectively, where X_2, is the chi-square goodness of fit statistic observed from fitting two dimensional model, X_3, is the chi-square goodness of fit statistic observed from fitting three dimensional model, while X_4, is the chi-square goodness of fit statistic observed from fitting four dimensional model.

From [1] final size epidemic data in table 1.2 and our discussion in section 6.3 on the computation of the degrees of freedom of the Pearson chi-square statistic, we know that if the two dimensional model is true then $X_2 \approx \chi^2_{13}$, if the three dimensional model is true then $X_3 \approx \chi^2_{12}$, while if the four dimensional model is true then $X_4 \approx \chi^2_{11}$.

Also, since the models are nested within each other, we can express their relationships in the form,

two dimensional model $\subseteq$ three dimensional model $\subseteq$ four dimensional model, such that the smaller model with fewer parameters has more degrees of freedom, while the larger models with more parameters has fewer degrees of freedom. Observe that we will have, $X_2 \geq X_3 \geq X_4$.

We now construct differences between the chi-square goodness of fit statistics from the two, three and four dimensional models, respectively as,

$D_{2,3} = X_2 - X_3 \geq 0$, $D_{2,4} = X_2 - X_4 \geq 0$ and $D_{3,4} = X_3 - X_4 \geq 0$.

If the two dimensional model is the sufficient fit to the final size epidemic data then, $D_{2,3} = X_2 - X_3 \approx \chi^2_1$. If the two dimensional model is not the better fit then, $D_{2,3} = X_2 - X_3 \gg \chi^2_1$ and the three dimensional model is a better fit on the final size epidemic data. Similarly, if the two dimensional model is a sufficient fit on the final size epidemic data then, $D_{2,4} = X_2 - X_4 \approx \chi^2_2$. If two dimensional model is not a sufficient fit on it on the final size epidemic data then, $X_2 - X_4 \gg \chi^2_2$ and four dimensional provides better fit to the final size

epidemic data. Also, if the three dimensional model is the sufficient model fit to the final size epidemic data then, $D_{3,4} = X_3 - X_4 \approx \chi_1^2$, if it is not the better model fit to the final size epidemic data then, $D_{3,4} = X_3 - X_4 \gg \chi_1^2$ and the four dimensional model provides the better model fit to the final size epidemic data.

7.3 Kolmogorov-Smirnov test.

Using the chi-square difference statistics $D_{2,3}$, $D_{2,4}$, and $D_{3,4}$, we plotted their empirical cumulative distribution functions with the corresponding cumulative distribution function of the hypothesized theoretical chi-square distribution and evaluated their test statistics for the three alternative hypotheses for the Kolmogorov-Smirnov test in [46], namely the two-sided test and the other two one-sided tests in section 6.5 conducted at the upper 1% and 5% significance.

Here, we have also adopted the notations of tail for the alternative hypotheses as in Mathworks documentation in section 6.5. The critical value is obtained from the Matlab function, kstest(dataset, 'CDF', cdf, 'Alpha',alpha, 'Tail', tail) in [46] at the given level of significance. Here, CDF, Alpha and tail are as defined in section 6.5. The test statistic T, the p and the critical values, the decision rules of the test $h = 0$ and $h = 1$ for not rejecting and for rejecting the null hypothesis for the three models are presented in sections 7.5, 7.7 and 7.10 respectively.

7.4 Proportion of the simulations rejected from the chi-square difference test.

Using the procedures in section 7.2, we investigated the properties of the models for the upper $\alpha = 5\%$ point of the chi-square distribution with 1 and 2 degree of freedoms given by the $1 - \alpha$ quantiles of the chi-square distributions, 3.841 and 5.991 respectively.

We reject the two dimensional model in favour of the three dimensional model, if $D_{2,3} >$ 3.841, when the true model is the two dimensional model. Also, we reject the two dimensional

in favour of the three dimensional model if $D_{2,4} > 5.991$ when the two dimensional model is true. We reject the three dimensional model in favour of the four dimensional model if $D_{3,4} > 3.841$ when the true model is the three dimensional model.

7.5 Chi-square difference and the Kolmogorov-Smirnov tests on the two dimensional final size epidemic data.

In order to implement the procedures discussed in section 7.2, we simulate two dimensional household final size epidemic data with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G, and large population size using the function, ThreeandfourandTwoSNsimhousesDifftwo and some subroutines as follow.

Run ThreeandfourandTwoSNsimhousesDifftwoks simulate household epidemic with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G. It calculates other corresponding parameters, computes the chi-square difference statistics, their mean and variance. It also computes the proportion of the simulations rejected from the chi-square difference test at 5% significance. These are accomplished using the subroutines in section 6.7

We implement these procedures with theoretical parameters corresponding to $z = 0.7298$ given by $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$, household structure in [1] but fifty times its population size, which is 70700 population size, minimum epidemic size of 1000 and simulation runs of 500 and obtained the following density histograms.

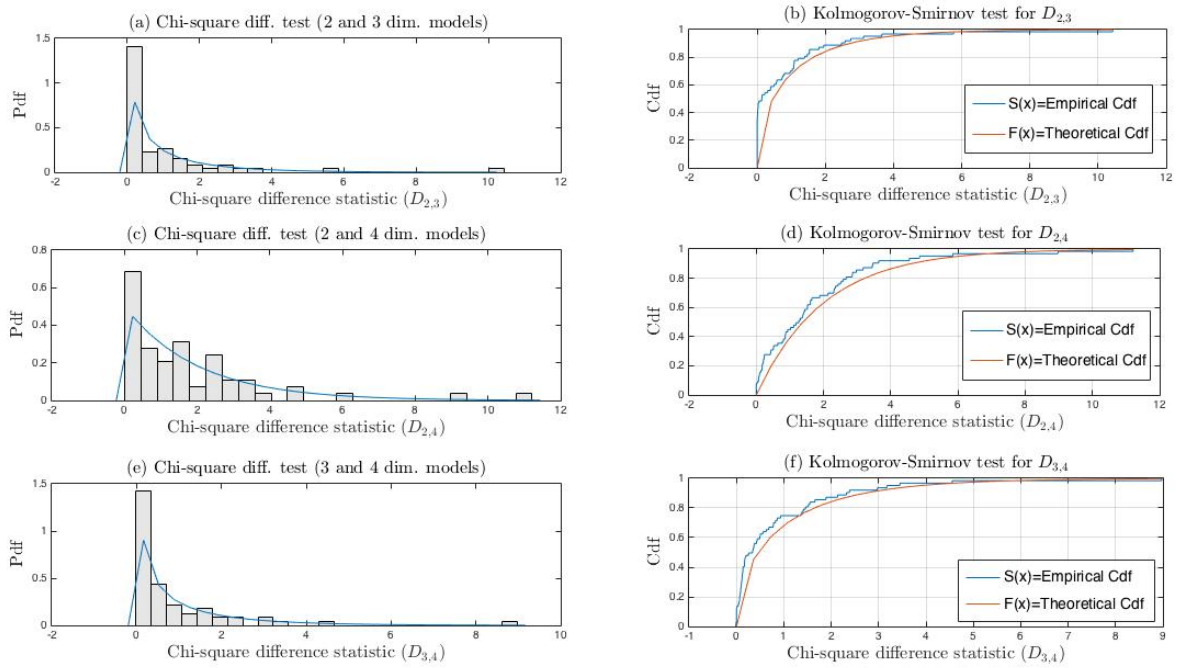

Figure 7.1: Density histograms of chi-square difference statistic to two dimensional final size epidemic data and plots of the empirical distribution of the chi-square difference statistic.

From figures 7.1 (b), (d) and (f), we see that the vertical distances between the theoretical and empirical distribution functions of $D_{3,4}$ is small compared to those of the $D_{2,3}$ and $D_{2,4}$. However the three models are sufficient fit to the final size epidemic data with the two dimensional model with two parameters most preferred model fit to the final size epidemic data.

7.6 Table of mean and variance of the chi-square difference tests on the two dimensional final size epidemic Data.

Chi-square difference statistic.	Simulated value		Theoretical Value.	
	mean	variance	mean	variance
$D_{2,3}$	0.5047	1.6876	1	2
$D_{2,4}$	1.3944	3.6819	2	4
$D_{3,4}$	0.8897	2.188	1	2

Table 7.1: Table of mean and variance of the chi-square difference tests on two dimensional final size epidemic data.

From table 7.1, we see that the mean and variance of the simulated chi-square difference statistic are approximately close to their theoretical counterparts having one or two degrees of freedom. For example $D_{2,3}$ have the mean 0.5047 and variance 1.6876, which is χ^2_1, $D_{2,4}$ have the mean 1.3944 and variance 3.6819 which is approximately χ^2_2, while $D_{3,4}$ have the mean 0.8897 and variance 2.188, which is approximately χ^2_1.

Difference Chi-square Statistic.	Upper 5% point	Proportion Rejected
$D_{2,3}$	3.841	0.0320
$D_{2,4}$	5.991	0.0160
$D_{3,4}$	3.841	0.0320

Table 7.2: Proportion of the simulations rejected from the chi-square difference test at 5% significance from the two dimensional epidemic data.

In table 7.2, the proportion of the simulations rejected are close to 0.05 as theoretically expected. This signifies that the three models fit well to the final size data.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, P=0.00000 T=0.347450	**h=1**, p=0.00000 T=0.347450	h=0, p=0.775693 T=0.015611
$D_{2,4}$	**h=1**, p=0.00000 T=0.162424	**h=1**, p=0.00000 T=0.162424	**h=1**, p=0.00000 T=0.020768
$D_{3,4}$	**h=1**, p=0.00000 T=0.137251	**h=1**, p=0.00000 T=0.137251	h=0, p=0.283982 T=0.035151

Table 7.3: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the two dimensional final size epidemic data in figure 7.1

In table 7.3, the null hypothesis for the two sided test from $D_{2,3}$ and $D_{3,4}$ are rejected owing the significant discrepancies between their cumulative distribution functions in one direction, while that of $D_{2,4}$ is rejected owing the significant discrepancies in both directions.

Thus $D_{2,3}$ with smaller difference in one direction between the cumulative distribution function is a better approximation. The three and four dimensional models are not significantly better that the two dimensional model.

7.7 Chi-square difference and the Kolmogorov-Smirnov tests on the three dimensional final size epidemic data.

We simulate household epidemics with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G, and large population size using the function, ThreeandTwoDimopton-ThreesimhousesDchsq as follows.

Run the function, ThreeandTwoDimoptonThreesimhousesDchsq with Gamma(a, b) infectious period distribution and theoretical parameters, λ_L, λ_G. It then calculates other corresponding parameters of the models, using Gamma(a, b) infectious period distribution. It computes the chi-square difference statistics, their mean, variance and the proportion of the simulations rejected from the chi-square difference test at 5% significance using the subroutines in section 7.5.

We implemented these procedures using the theoretical parameters in section 7.2 for $\varepsilon = 0.1, 0.3$ and plotted the density histograms of the chi-square difference statistics superimposed with their theoretical chi-square distribution. Including those of the empirical distribution functions with the cumulative distribution function of the hypothesized distribution function.

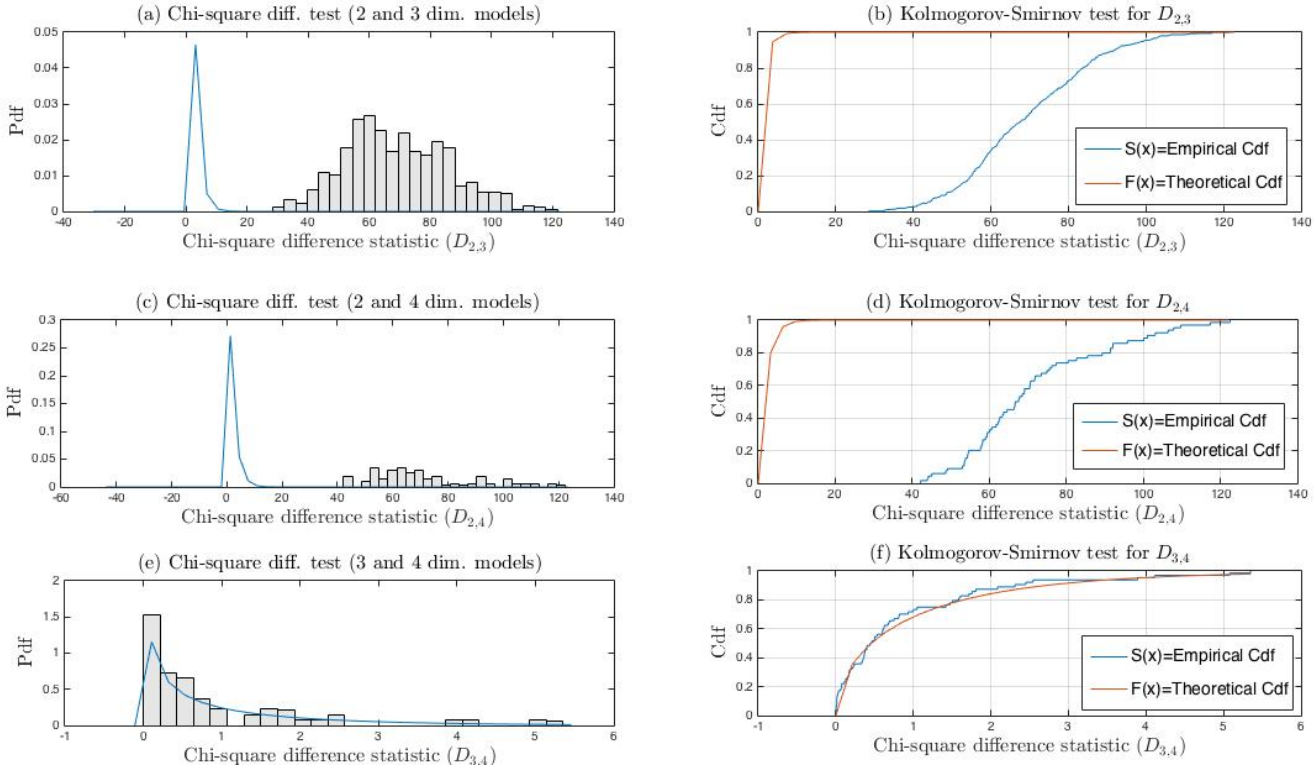

Figure 7.2: Density histograms of the chi-square difference statistic on the three dimensional final size epidemic data and those of the empirical and cumulative distribution functions when the misclassification probability, $\varepsilon = 0.1$.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, P=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{2,4}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{3,4}$	**h=1**, p=0.000316, T=0.093149	**h=1**, p=0.000158 T=0.093149	h=0, p=0.697474 T=0.018654

Table 7.4: Table of summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the three dimensional final size epidemic data when $\varepsilon = 0.1$ in figures 7.2 (b), (d) and (f).

In table 7.4, we see that the null hypothesis for the two sided test is rejected from the three statistics owing to the significant differences between their empirical cumulative functions and the cumulative of the chi-square distribution functions.

The difference between the empirical cumulative of $D_{3,4}$ and the cumulative of the hypothesized distribution functions in one direction is small. The three and four dimensional models are the significantly better than the two dimensional model. The four dimensional model is not significantly better than the three dimensional model.

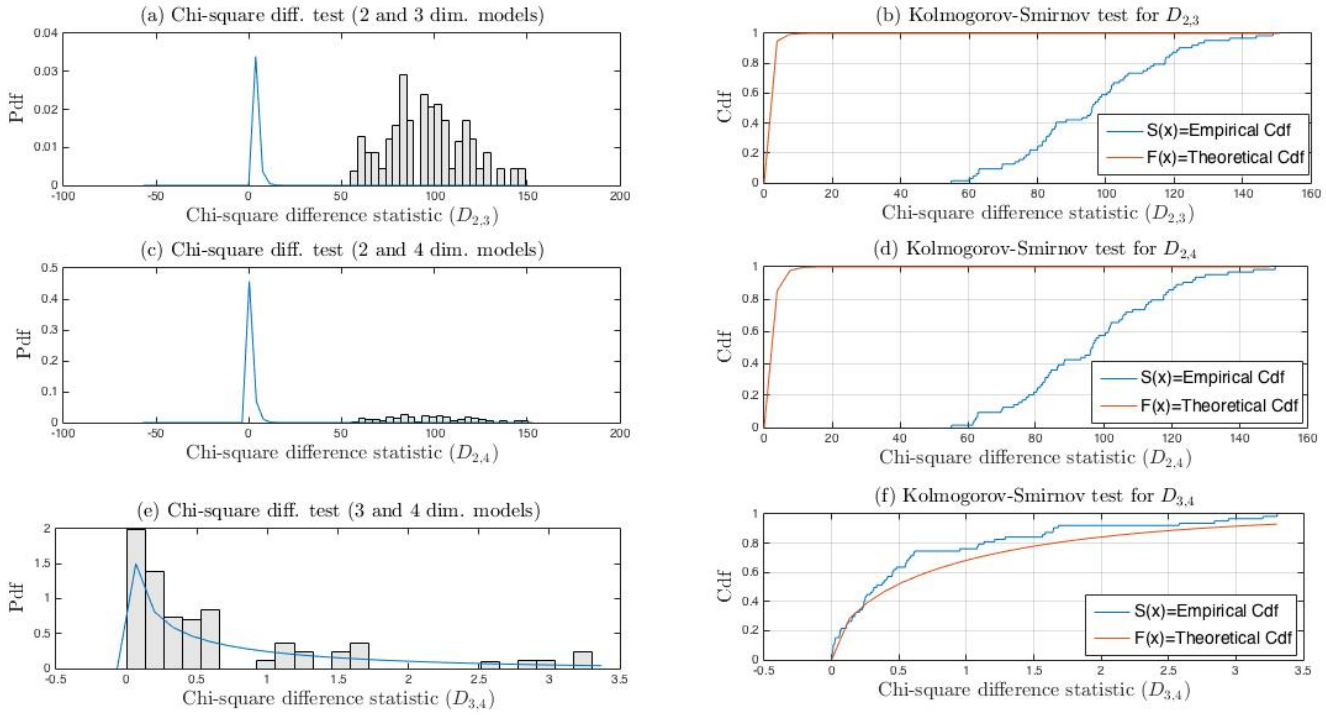

Figure 7.3: Density histograms of the chi-square difference statistic on the three dimensional final size epidemic data and those of the empirical and cumulative distribution functions, when the misclassification probability, $\varepsilon = 0.3$.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$.	**h=1**, P=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{2,4}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{3,4}$	**h=1**, p=0.00000 T=0.179102	**h=1**, p=0.00000 T=0.179102	h=0, p=0.148752 T=0.043320

Table 7.5: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the three dimensional final size epidemic data when $\varepsilon = 0.3$ in figures 7.3 (b), (d) and (f).

In tables 7.5 and 7.4 present similar behaviours. The three and four dimensional models are appropriate fit to the data.

Difference Chi-square Statistic.	Upper 5% point	Proportion Rejected.		
		$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 0.3$
$D_{2,3}$	3.841	0.0160	1	1
$D_{2,4}$	5.991	0.0480	1	1
$D_{3,4}$	3.841	0.0780	0.0540	0

Table 7.6: Proportion of the simulations rejected from the chi-square difference test at 5% significance from the three dimensional final size epidemic data.

In table 7.6, we see that with increasing misclassification probabilities in the permissible region the three and four dimensional models fit significantly better than the two dimensional model.

7.8 Table of mean and variance of the chi-square difference statistic on the three dimensional final size epidemic Data.

From table 7.7, we see that, the mean and variance of the simulated chi-square difference statistic, $D_{2,3}$, $D_{2,4}$, $D_{3,4}$ are approximated close to their theoretical counterparts with 1, 2 and 1 degrees of freedom respectively and with large values of the misclassification proba-

Chi-sq. diff. stat.	$\varepsilon = 0.0$		$\varepsilon = 0.1$		$\varepsilon = 0.3$		Theor. Value	
	mean	var	mean	var	mean	var	mean	var
$D_{2,3}$	0.4340	1.0322	69.07	284.6	93.005	356.71	1	2
$D_{2,4}$	1.3136	3.4747	70.086	287.41	93.945	354.59	2	4
$D_{3,4}$	0.8795	2.3852	1.0162	2.5045	0.9400	1.3059	1	2

Table 7.7: The mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data simulated with misclassification probabilities, $\varepsilon = 0.0$, 0.1, 0.3. Here, Theor., is the theoretical mean or variance.

bilities, only those of $D_{3,4}$ remains consistent, while those of $D_{2,3}$ and $D_{2,4}$ increases with increasing misclassification probabilities in the final size epidemic data.

7.9 Plots of the mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data.

We complement our studies in section 7.7 by simulating household epidemic with Gamma(a, b) infectious period distribution and theoretical parameters, λ_L and λ_G using the function, ThreefourTwoDimplotschsqlik in section 6.9 with subroutines in subsection 6.7.

Employing these procedures with theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$ for a range of $\varepsilon \in [0, 0.5)$ we simulate household epidemic, estimate the parameters of the three models and compute their chi-square difference statistics, their mean and variance and also plot the mean of the chi-square difference statistic for $\varepsilon[0, 0.1]$ those of the proportion of the simulations rejected from the Pearson chi-square test, at 5% significance.

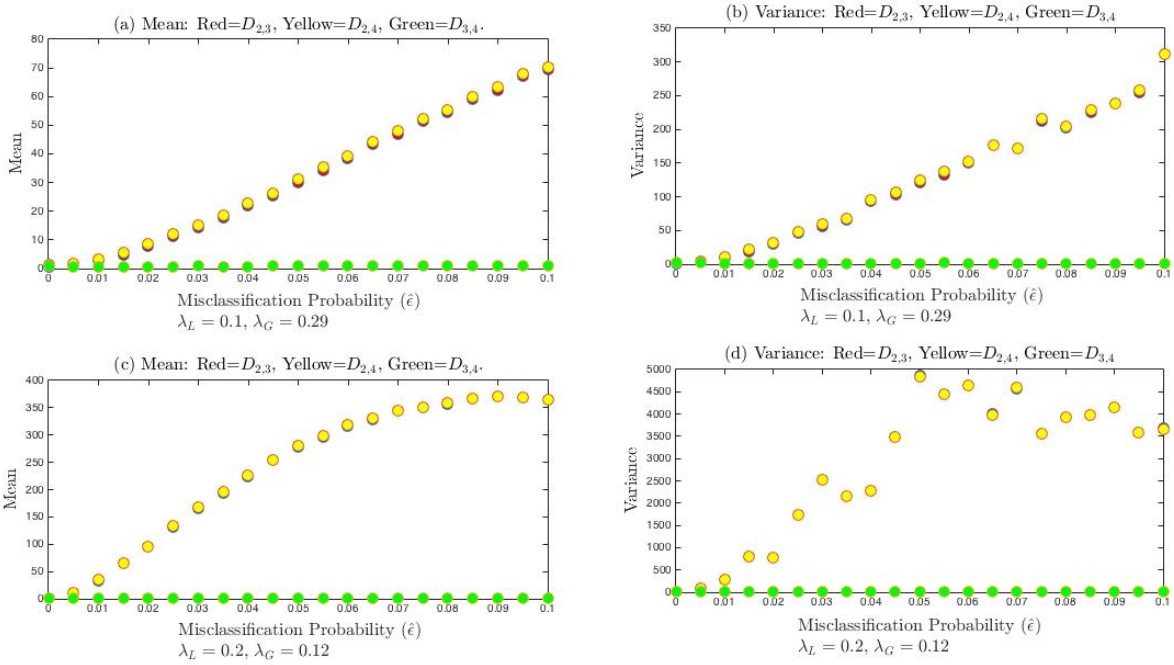

Figure 7.4: The mean and variance of the chi-square difference statistic on the three dimensional final size epidemic data for $\varepsilon \in [0, 0.1]$, step size of 0.005.

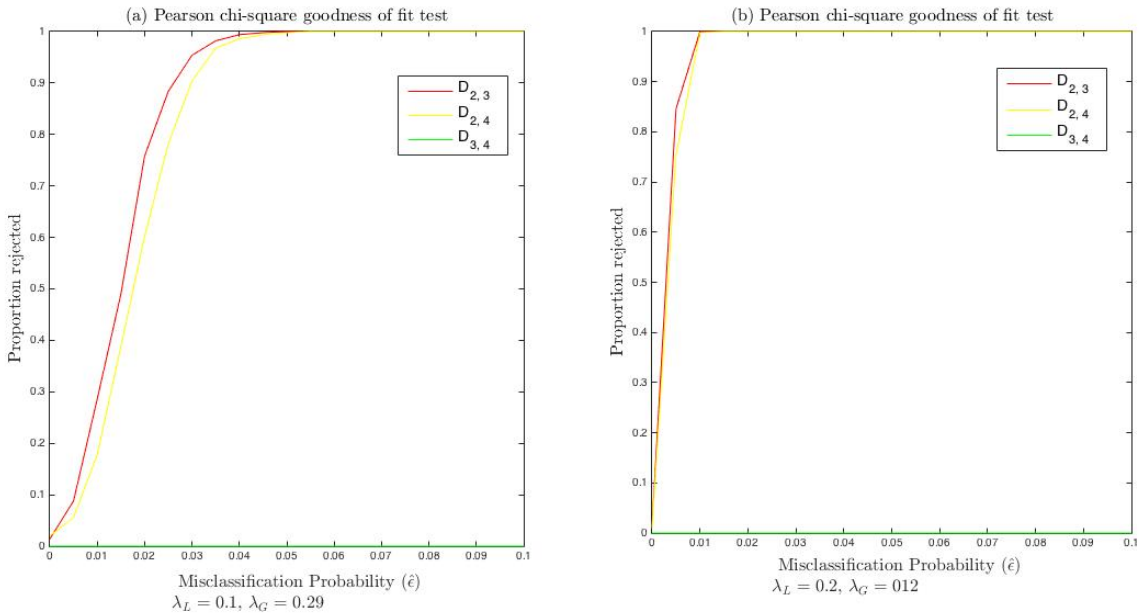

Figure 7.5: Proportion of the simulations rejected at 5% significance from the chi-square difference test for $z = 0.7298$ and $z = 0.2144$ when it is the three dimensional final size epidemic data.

We see from figure 7.5 that the proportion of the simulations rejected from the chi-square difference tests for $D_{2,3}$ and $D_{2,4}$ when the theoretical parameters corresponds to $z = 0.7298, 0.2144$, increases towards 1, while those of $D_{3,4}$ reduces to 0 at the upper 5% points. These values are significantly different from the required proportion of the simulations rejected at 5% significance when the null hypothesis is true.

The three dimensional model is significantly better than the two dimensional model, while the four dimensional model is better than the three dimensional model when the misclassification probability is large.

7.10 The chi-square difference and Kolmogorov-Smirnov tests on the four dimensional final size epidemic data.

Using the misclassification probabilities in table 6.9, we simulate household epidemic with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G with the function FourDimThreeATwoSNsimhousesDchsq and as follows.

Run FourDimThreeATwoSNsimhousesDchsq to simulate two dimensional household epidemic with Gamma(a, b) infectious period distribution, theoretical parameters, λ_L, λ_G. It then calculates corresponding parameters. It computes the chi-square difference statistics and plot their density histogram for the three model. It compute the mean and variance of the chi-square difference statistic and the proportion of the simulations rejected from the chi-square difference test at 5% significance.

We implement these procedures with the theoretical parameters $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$, household structure in [1] and population size of 70700, minimum epidemic size of 1000. The chi-square difference statitsics of the three models, their empirical cumulative distribution function superimposed with their theoretical counterparts are then obtained as follows.

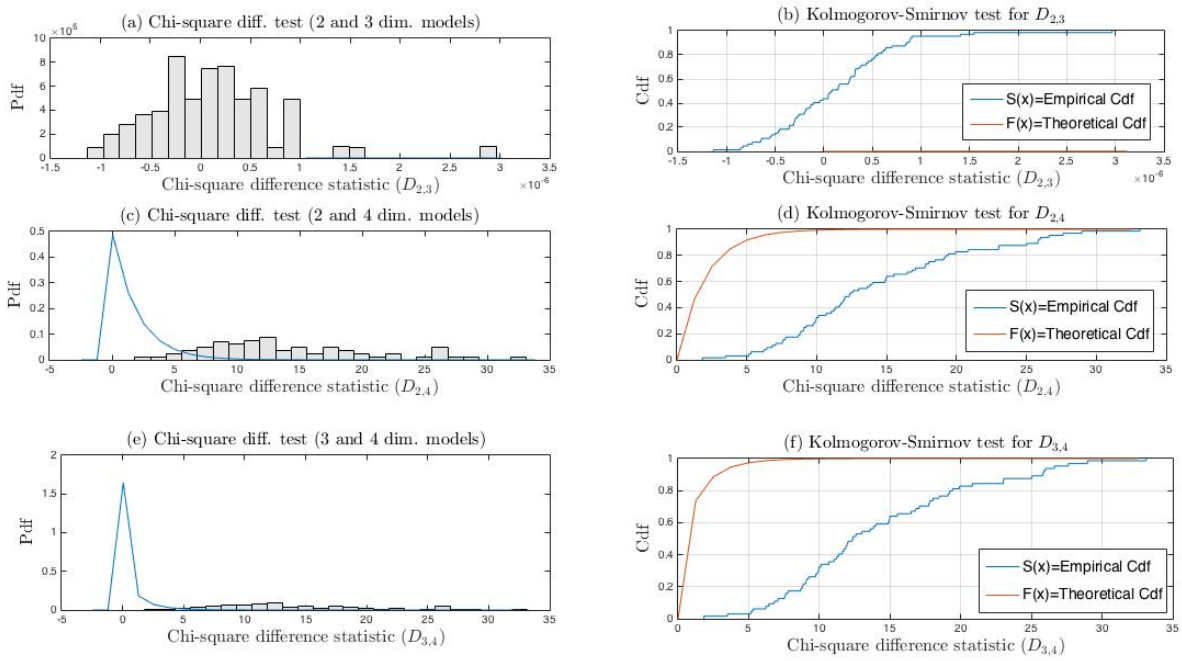

Figure 7.6: Density histograms of chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when $\varepsilon_{FN} = 0$ and $\varepsilon_{FP} = 0.2$.

In figure 7.6 (a)-(f) we see wide discrepancies between the empirical cumulative of $D_{2,3}$, $D_{2,4}$ and $D_{3,4}$ and the cumulative of the hypothesized chi-square distributions.

The two and three dimensional models failed to fit the four dimensional final size epidemic data when the misclassification probability is large and far apart from each other. The four dimensional model is significantly better than the two and three dimensional models.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, P=0.00000 T=0.998626	**h=1**, p=0.00000 T=0.998626	h=0, p=1.00000 T=0.00000
$D_{2,4}$	**h=1**, p=0.00000 T=0.893736	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=0.893736
$D_{3,4}$	**h=1**, p=0.00000 T=0.946715	=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=0.946715

Table 7.8: Summary of the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0.2$ in figure 7.6(b), (d) and (f).

From table 7.8, the null hypothesis from the two sided test in the three cases of $D_{2,3}, D_{2,4}, D_{3,4}$ are rejected at the 0.05 significance. The empirical cumulative distribution functions from the three cases are not sufficient approximations to the cumulative of the hypothesized chi-square distribution functions.

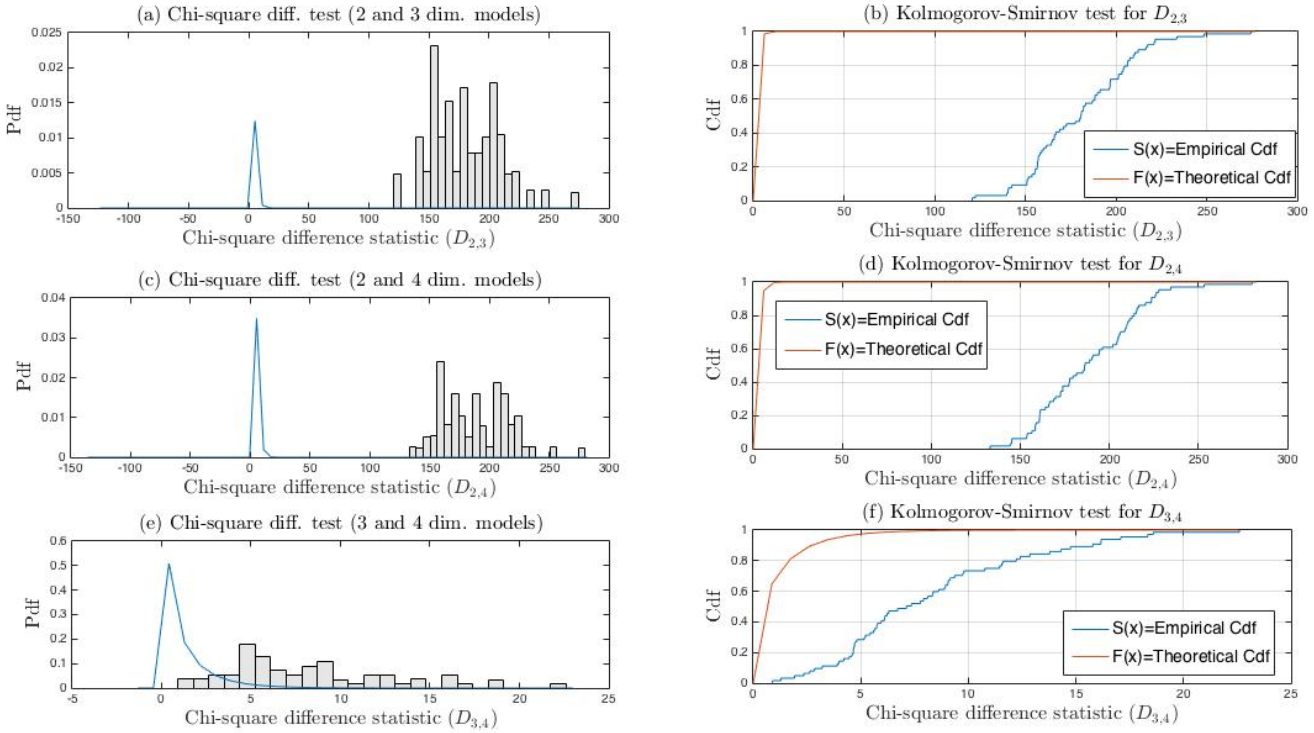

Figure 7.7: Density histograms of chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0$.

In figure 7.7 (a)-(f), similar behaviours in figure 7.6 (a)-(f) are presented. The four dimensional model is significantly better than the two and three dimensional models.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{2,4}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{3,4}$	**h=1**, p=0.00000 T=0.840152	h=0, p=0.999998 T=0.000002	**h=1**, p=0.00000 T=0.840152

Table 7.9: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0$ in figure 7.7 (b), (d) and (f).

In table 7.9, similar behaviours in table 7.8 are observed. The four dimensional model is significantly better than the two and three dimensional models.

In figures 7.8 (a)-(f), we see significant vertical distances between the theoretical and empirical distribution functions of the chi-square different statistic from the two and three dimensional models are observed, while the vertical distance between the two cumulative distribution functions from the four dimensional model is small.

This shows that the four dimensional model is significantly better then the two and three dimensional models.

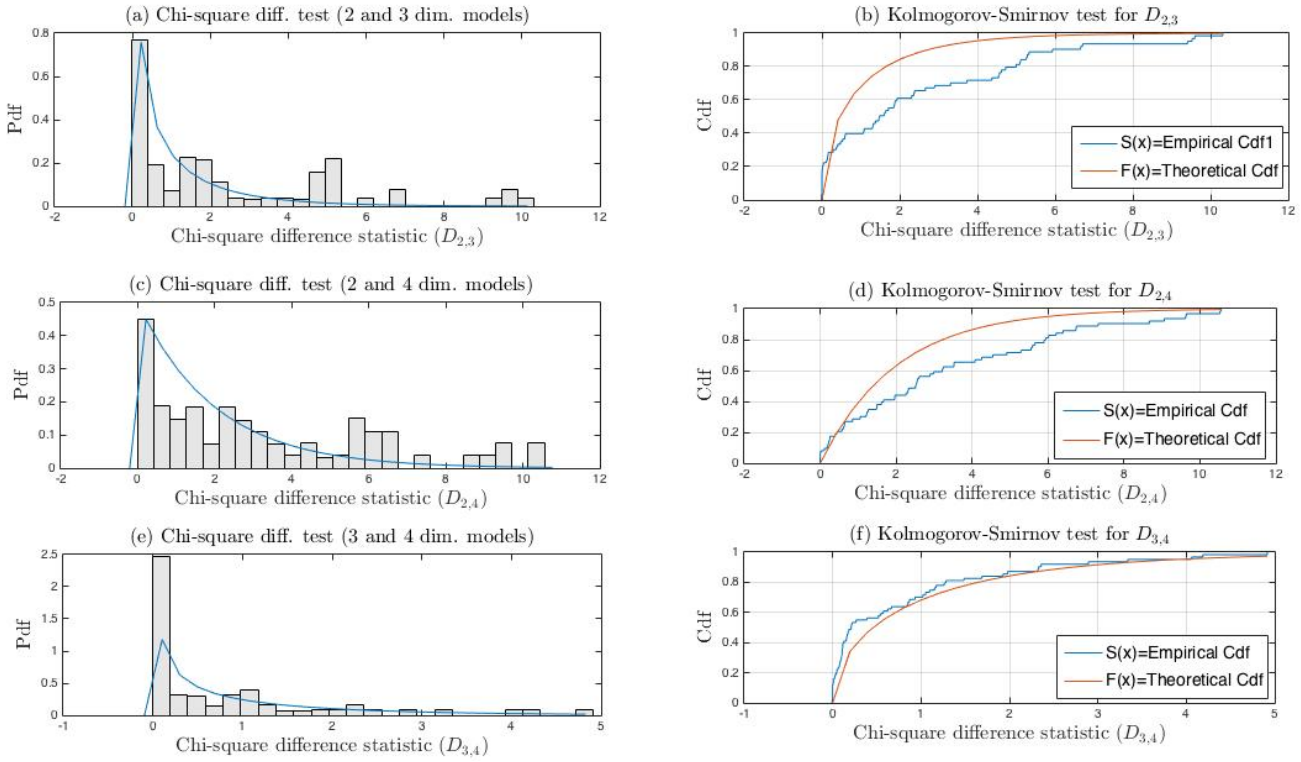

Figure 7.8: Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.01$ and $\varepsilon_{FP} = 0.02$.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, P=0.00000 T=0.318170	**h=1**, p=0.00000 T=0.159023	**h=1**, p=0.00000 T=0.318170
$D_{2,4}$	**h=1**, p=0.00000 T=0.236886	**h=1**, p=0.005298 T=0.072033	**h=1**, p=0.00000 T=0.236886
$D_{3,4}$	**h=1**, p=0.00000 T=0.172568	**h=1**, p=0.00000 T=0.172568	h=0, p=0.980825 T=0.004079

Table 7.10: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.01$, $\varepsilon_{FP} = 0.02$ in figure 7.8 (b), (d) and (f).

In table 7.10, the null hypothesis is rejected for the two sided test from $D_{2,3}, D_{2,4}$ owing to the significant differences between their empirical cumulative distribution functions and the cumulative of the chi-square distribution in both directions, while the difference in the case of $D_{3,4}$ occurred in one direction.

The empirical cumulative distribution function from $D_{3,4}$ is a better approximation of the cumulative of the chi-square distribution function. The three and four dimensional models are significantly better than the two dimensional model.

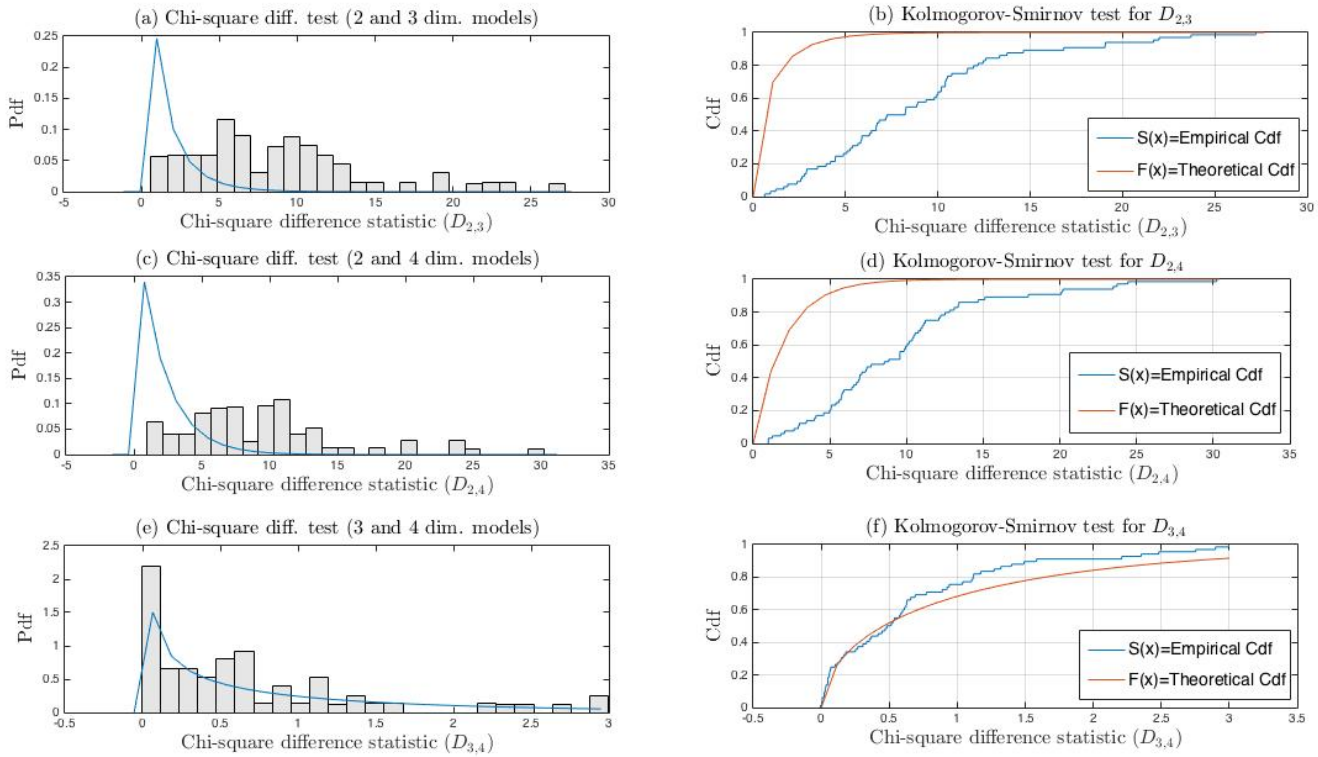

Figure 7.9: Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.02$ and $\varepsilon_{FP} = 0.01$.

From figures 7.9 (a)-(f) , we see similar behaviours to figure 7.8 (a)-(f).

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, p=0.00000 T=0.800637	h=0, p=0.00000 T=1.00000	**h=1**, p=0.00000 T=0.800637
$D_{2,4}$	**h=1**, p=0.00000 T=0.730849	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=0.730849
$D_{3,4}$	**h=1**, p=0.000001 T=0.120286	**h=1**, p=0.00000 T=0.120286	**h=1**, p=0.045004 T=0.055348

Table 7.11: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$ in figures 7.9 (b), (d) and (f).

In table 7.11, while the null hypothesis for the two sided test from $D_{2,3}$ and $D_{2,4}$ are rejected owing to the significant differences between their empirical cumulative distribution functions and the cumulative of the chi-square distribution function in one direction, that from $D_{3,4}$ is rejected owing the significant differences in both directions. We see that the four dimensional model is significantly better than the two and three dimensional models.

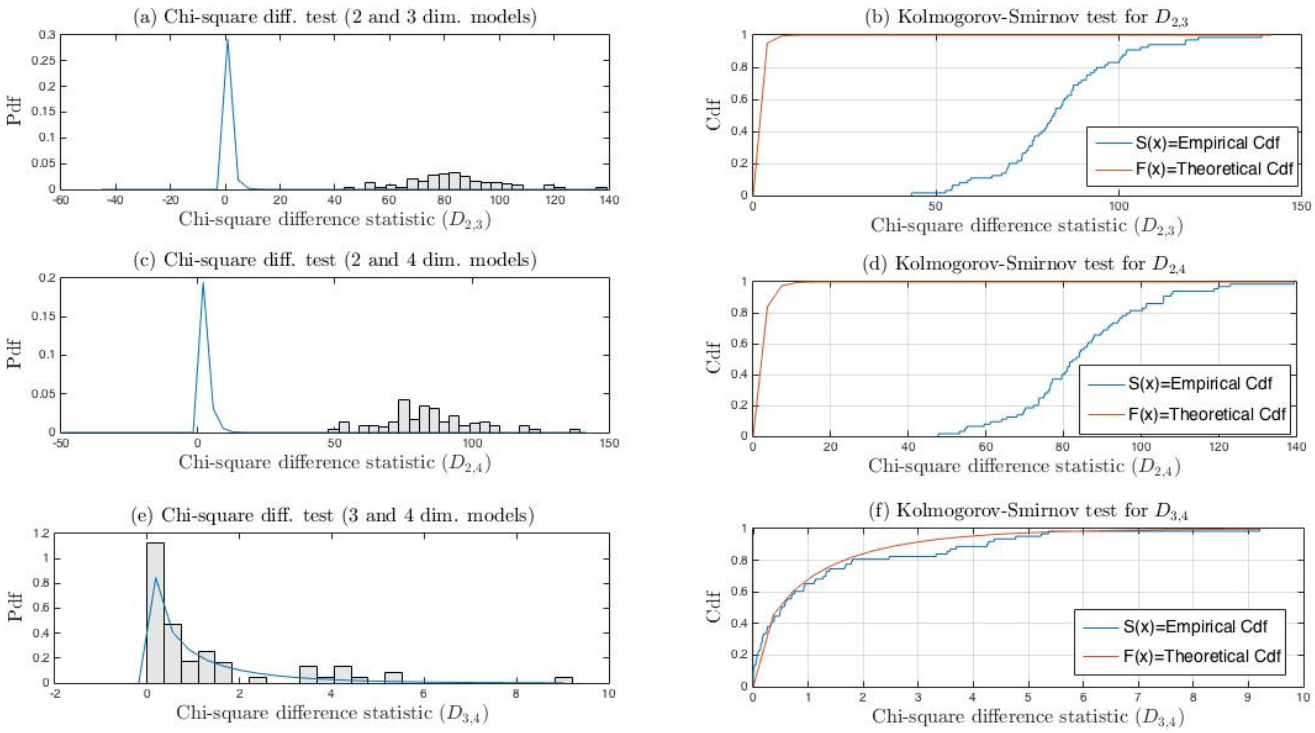

Figure 7.10: Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.2$ and $\varepsilon_{FP} = 0.3$.

In figures 7.10 (a), (d) and (f), we see that the empirical cumulative distribution functions from $D_{2,3}$ and $D_{2,4}$ are not good approximations of the cumulative of the hypothesized chi-square distribution functions unlike that of $D_{3,4}$ which is close to the cumulative of the hypothesized chi-square distribution function.

The three and four dimensional model are significantly better than the two dimensional model.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S 1	Tail (F<S) -1
$D_{2,3}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{2,4}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{3,4}$	**h=1**, p=0.000016 T=0.107749	**h=1**, p=0.002281 T=0.077626	**h=1**, p=0.000008 T=0.107749

Table 7.12: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.3$ in figures 7.10 (b), (d) and (f).

In table 7.12, the two sided test rejects the null hypothesis at 0.05 significance level, for the three cases. These decision outcomes are in agreement with the significant differences between the empirical cumulative distribution functions and the cumulative of the chi-square distribution functions in figures 7.10 (b), (d) and (f). The three and four dimensional model are significantly better than the two dimensional model.

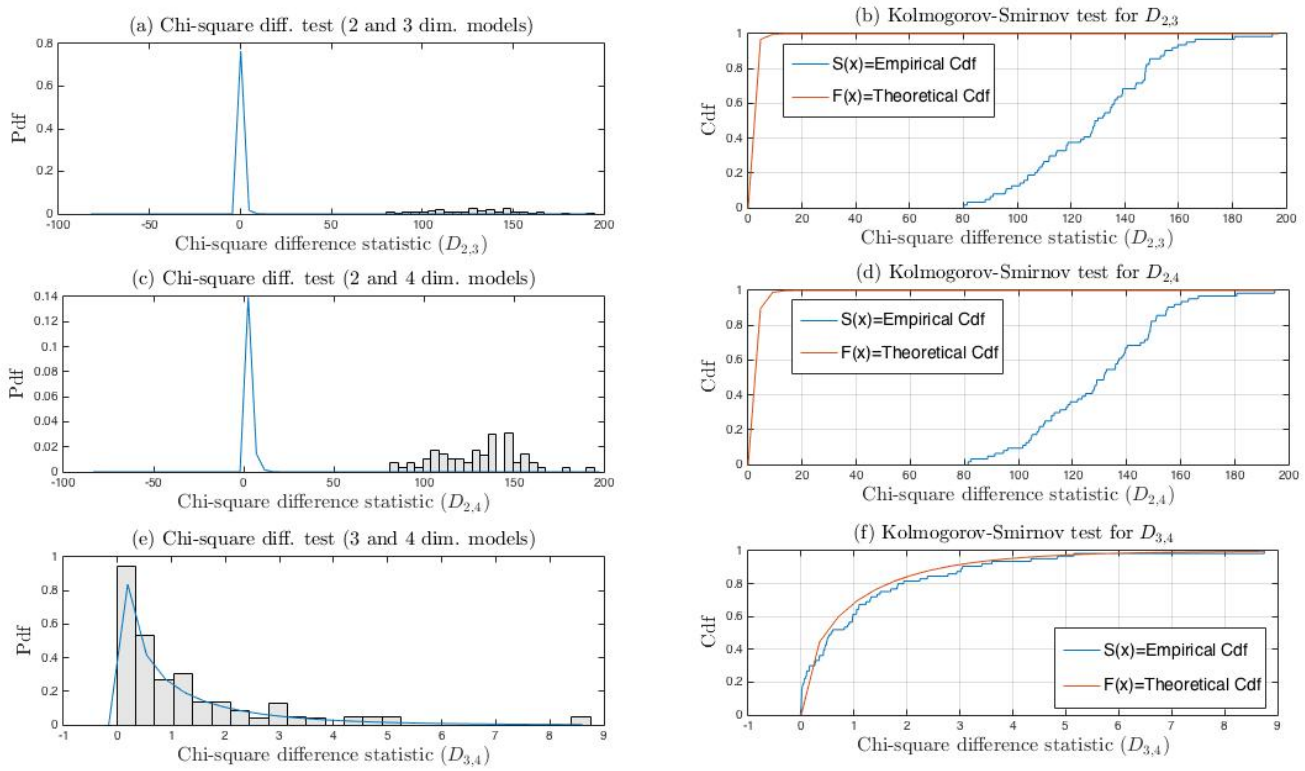

Figure 7.11: Density histograms of the chi-square difference statistic on the four dimensional final size epidemic data superimposed with their theoretical counterparts and those of the empirical cumulative distribution functions with their theoretical counterparts when, $\varepsilon_{FN} = 0.3$ and $\varepsilon_{FP} = 0.2$.

In figures 7.11 (a)-(f), similar behaviour in figures 7.10 (a)-(f) are observed.

Ch. Diff. Stat.	Tail (F>S and F<S) 0	Tail (F>S) 1	Tail (F<S) -1
$D_{2,3}$	**h** = 1, P=0.00000 T=1.00000	**h=0**, p= 1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{2,4}$	**h=1**, p=0.00000 T=1.00000	h=0, p=1.00000 T=0.00000	**h=1**, p=0.00000 T=1.00000
$D_{3,4}$	**h=1**, p=0.00000 T=0.120183	**h=1**, p=0.004513 T=0.073130	**h=1**, p=0.000000 T=0.120183

Table 7.13: Summary from the Kolmogorov-Smirnov goodness of fit tests with the upper 5% points from the four dimensional final size epidemic data when $\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$ in figures 7.11 (b), (d) and (f).

In table 7.11, we see similar decision outcomes in table 7.12. The four dimensional model is significantly better than the two and three dimensional models.

7.11 Table of mean and variance of the chi-square difference statistic.

Miscl. Prob.	$D_{2,3}$		$D_{2,4}$		$D_{3,4}$	
	mean	variance	mean	variance	mean	variance
$\varepsilon_{FN} = 0.0$, $\varepsilon_{FP} = 0.0$	0.46581	1.0916	1.2936	2.8806	0.82778	1.8794
$\varepsilon_{FN} = 0$, $\varepsilon_{FP} = 0.2$	1.54E-08	3.89E-13	14.986	59.37	14.986	59.37
$\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0$	181.27	851.5	189.87	800.57	8.5973	24.226
$\varepsilon_{FN} = 0.01$, $\varepsilon_{FP} = 0.02$	2.4539	8.7633	3.277	10.781	0.82313	2.2408
$\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.01$,	8.8483	34.703	9.4906	36.606	0.64231	0.71849
$\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.3$	85.34	311.93	86.774	321.26	1.4343	2.9258
$\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$	130.47	461.77	131.77	451.15	1.2955	2.6945

Table 7.14: The mean and variance of the chi-square difference statistic on the four dimensional final size epidemic data simulated with misclassification probabilities in table 6.9. Here miscl. Prob. are the misclassification probabilities. The theoretical mean and variance of the chi-square difference statistics are as defined in table 7.7.

In table 7.14 the effects of the misclassification probabilities on the mean and variance of the chi-square difference statistics is studied. Here, we see that if the misclassification

probabilities are close to 0, then the mean and variance of the chi-square difference statistics sufficiently approximate the theoretical mean and variance of the hypothesized chi-square distribution for the three cases on four dimensional final size data. If the misclassification probabilities are far apart from each other then disproportionate mean and variance of the chi-square difference statistics are obtained as shown in table 7.14.

7.12 Plots of the mean and variance of the chi-square difference statistic on the four dimensional final size epidemic data.

We explored the estimates of the parameters along the diagonal of the the misclassification probabilities region with theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$ and $\varepsilon_{FP} \in [0, 0.2]$.

This involves simulating four dimensional household epidemic along the line $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$ with Gamma(a, b) infectious period distribution using the function, FourThreeTwoDonFourfposChsqlik in section 6.13.

It calculates other corresponding parameters of the three models and computes the chi-square difference statistics from the three models. It plot the mean and variance and the proportion of the simulations rejected from the chi-square difference test at 5% significance.

These are accomplished using the subroutines in section 7.10 and demonstrated in the figures 7.12 (a)-(d) and figures 7.13 (a) and (b).

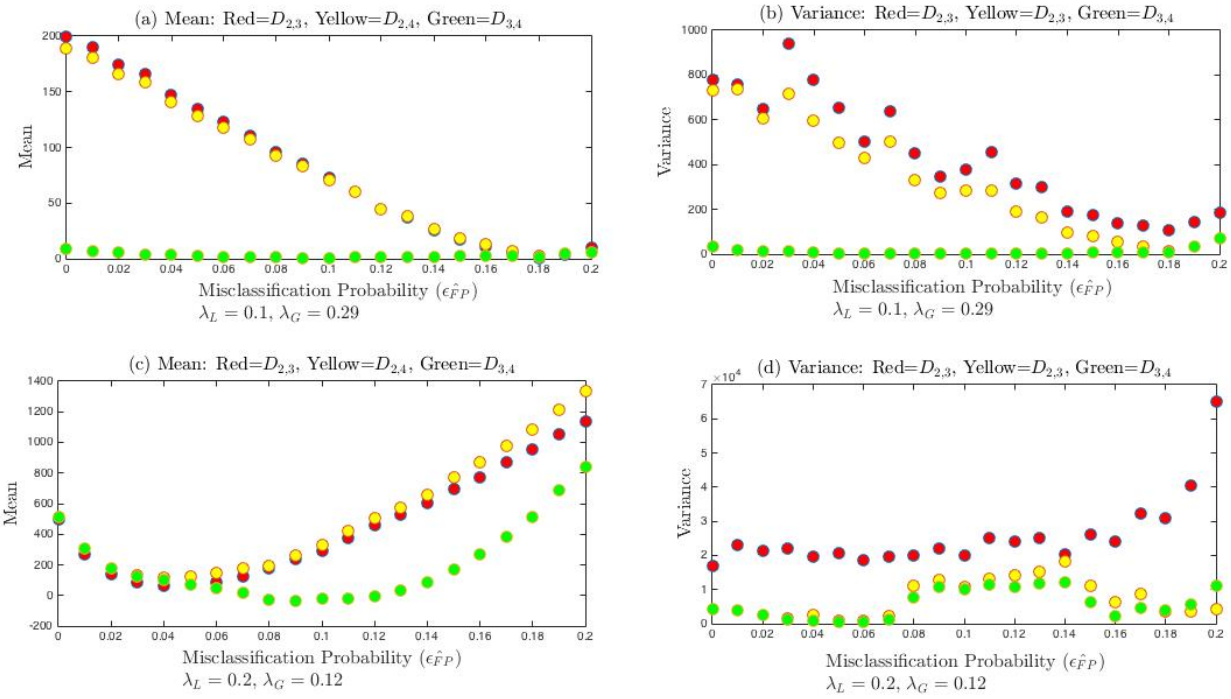

Figure 7.12: Plots of the mean and variance of the chi-square difference statistic for the three models, explored along the diagonal, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$, with step size of 0.01 and theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$ respectively.

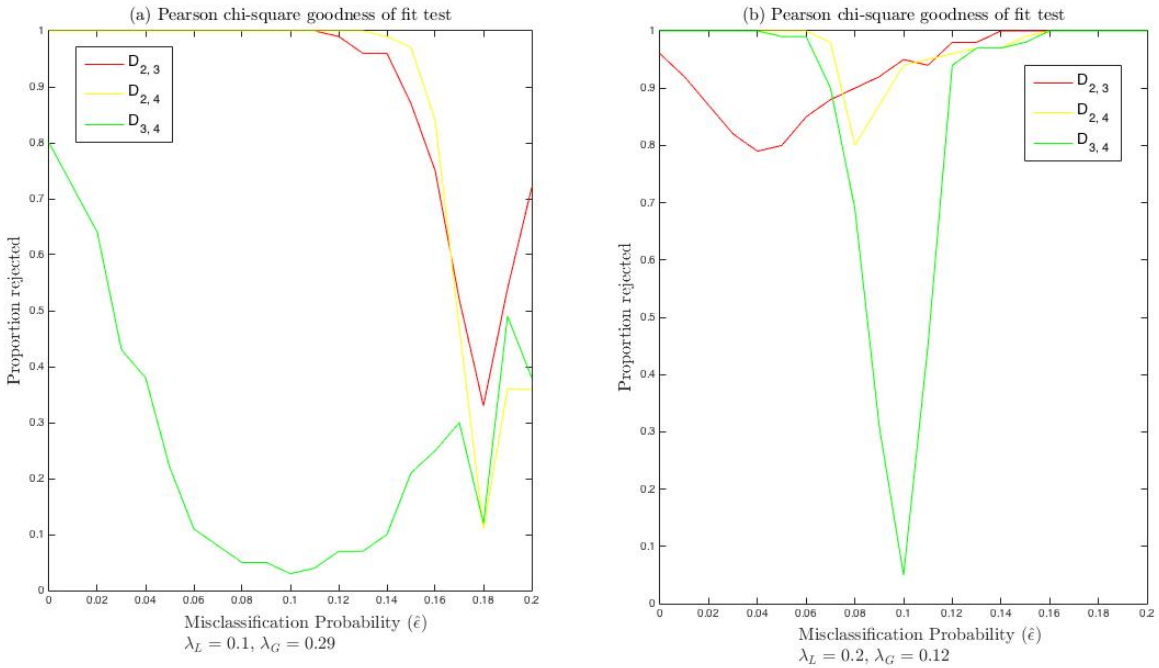

Figure 7.13: Proportion of the simulations rejected at 5% significance from the chi-square difference test with theoretical parameters corresponding to $z = 0.7298$ and $z = 0.2144$ when the true data is four dimensional final size epidemic data.

From figure 7.13 (a), we see that when $\varepsilon_{FP} \geq 0.15$ and the theoretical parameters corresponds to $z = 0.7298$, the proportion of the simulations rejected from the chi-square difference tests for $D_{2,3}$ and $D_{2,4}$ is consistently 1 and decreases when $\varepsilon_{FP} \geq 0.15$. While that of $D_{3,4}$ decreases for $\varepsilon_{FP} < 0.1$ and then gradually increases when $\varepsilon_{FP} > 0.1$.

In figure 7.13 (b) $D_{3,4}$ has the best proportion rejected especially when $\varepsilon = 0.1$, since the four dimensional model reduces to the three dimensional for this value of ε. Thus, the three dimensional model has the best proportion rejected at $\varepsilon = 0.1$, while the two dimensional model failed fitting the four dimensional final size epidemic data as expected.

In general, the three and four dimensional models sufficiently fit the final size epidemic data.

7.13 Fitting the three models to [1] Tecumseh Michigan Influenza A(H3N2) epidemic data using chi-square difference statistic.

We fitted the three models to [1] Tecumseh Michigan Influenza A(H3N2) epidemic data, the $1975-1976$ B(H1N1) and $1978-1979$ A(H1N1) Seattle influenza datasets with the chi-square difference test using the function, Addychsdiff with Gamma(a, b) infectious period distribution and theoretical parameters, λ_L, and λ_G describe as follows.

Run the function, Addychsdiff to estimate the parameters of the three models with $T_I =$ Gamma$(2, 2.05)$ infectious period distribution and compute the chi-square difference statistic for the three models using subroutines in section 7.10 as in table 7.15 as follow.

Chi-sq. Diff. and P-values.	$D_{2,3}$ value	$D_{2,4}$ value	$D_{3,4}$ value
Ch-sq. Diff.	0	0	0
P-values	P ≈ 1	P ≈ 1	P ≈ 1

Table 7.15: Table of the chi-square difference statistic for the three models with their corresponding P-values for the tests.

7.14 Fitting the three models to [28] Seattle Influenza epidemic data using chi-square difference statistic.

Par. of the infect. Per. dist.	$D_{2,3}$ value	$D_{2,4}$ value	$D_{3,4}$ value
a=1, b=4.1	0	0.2619	0.2619
P-value	P$\approx$ 1	P =0.8773	P=0.6088
a=2, b=4.1/2	0	2.2480	2.2480
P-value	P$\approx$ 1	P=0.3250	P=0.1338
a=5, b=4.1/5	0	0.8627	0.8627
P-value	P$\approx$ 1	P= 0.6496	P= 0.3530

Table 7.16: Table of chi-square difference statistic for the three models from the Seattle 1975−1976 B(H1N1) influenza epidemic with T_I = Gamma(a, b) infectious period distribution.

Par. of the infect. Per. dist.	$D_{2,3}$ value	$D_{2,4}$ value	$D_{3,4}$ value
a=1, b=4.1	0	0	0
P-value	P$\approx$ 1	P$\approx$ 1	P$\approx$ 1
a=2, b=4.1/2	0	0.001	0.001
P-value	P$\approx$ 1	P= 0.9995	P= 0.9748
a=5, b=4.1/5	0	0.0129	0.0129
P-value	P$\approx$ 1	P= 0.9936	P=0.9096

Table 7.17: Table of chi-square difference statistic for the three models from the Seattle 1978−1979 A(H1N1) influenza epidemic with T_I = Gamma(a, b) infectious period distribution

7.15 Discussion

We found that with misclassification error in the final size data, it becomes difficult to use the two dimensional model to fit the final size data as discussed in sections 5.8.1, 5.6, and 6.11.

The preferred model fit is that whose estimates are obtained taking these percentage errors into consideration. These behaviours are captured in section 7.7, in which only the three and four dimensional models fit the three dimensional final size data sufficiently well, when the misclassification probabilities are not close to 0.

These behaviours are further corroborated in table 7.7, in which only the mean and variance of the chi-square difference statistics $D_{3,4}$ are asymptotic to χ_1^2 as the misclassification probabilities is varied. We see that the mean and variance of $D_{2,3}$ and $D_{2,4}$ increase with increasing misclassification probabilities as shown in table 7.7.

This signifies that $D_{2,3} \gg \chi_1^2$ and $D_{2,4} \gg \chi_2^2$ and therefore the three and four dimensional models are significantly better than the two dimensional model. These behaviours are observed in figures, 7.2, 7.3 and 7.4.

With sufficiently large values of the misclassification probabilities away from 0 only $D_{3,4}$ is approximately χ_1^2, while $D_{2,3} \gg \chi_1^2$ and $D_{2,4} \gg \chi_2^2$. In other words only the three and four dimensional models are significantly better than the two dimensional model. From figures 6.6 and 6.7, we see that the four dimensional model is significantly better under this circumstance as the three dimensional model struggled fitting to the final size data as discussed in table 7.14.

Also, from table 7.15, 7.16 and 7.17, the observed chi-square difference test statistics are insignificant for the [1] Tecumseh Michigan influenza epidemic data and [28] Seattle influenza epidemic data respectively. The tests are therefore insignificant and the three and four dimensional models are not better than the two dimensional model.

Chapter 8

Estimation in the presence of model misspecification.

8.1 Introduction

If the model is estimated using a different infectious period distribution from that used for the simulations, then how does this affect the precision of the estimates? This is the focus of our studies in this section.

This is a misspecification problem which may sometimes be taken for misclassification of the epidemic data. It is therefore necessary to study these scenarios using simulations in order to understand their effects on the estimates of the parameters and be able to differentiate it from misclassification of the epidemic data.

We do this with large population size and theoretical parameters which give global infection in our simulations. We have therefore employed the theoretical parameters $\lambda_L = 0.1$, $\lambda_G = 0.29$ used in our previous studies to achieve this.

This chapter is organised in the following form.

In section 8.2, we simulate two dimensional household epidemic with exp(4.1) infectious period distribution and estimated the model with Gamma$(2, 4.1/2)$ infectious period distribution while in section 8.3 we simulate two dimensional household epidemic with Gamma$(2, 4.1/2)$ infectious period distribution and estimate the model with exp(4.1) infectious period distri-

bution. In section 8.4 we discussed the results of our studies from sections 8.2 and 8.3.

In section 8.5, we studied the effects of model misspecification on the estimates of the three models by simulating two dimensional household epidemic with exp(4.1) infectious period distribution and estimating the models with Gamma$(2, 4.1/2)$ infectious period distribution. While in section 8.6, we simulate two dimensional household epidemic with Gamma$(2, 4.1/2)$ infectious period distribution and estimate the models with exp(4.1) infectious period distribution. In section 8.7, we discussed the behaviours of the three models on the two dimensional epidemic data in the face of misspecification.

In section 8.9, we studied the effects of misspecification on the model estimates in the face of misclassification by simulating three dimensional household epidemic with exp(4.1) infectious period distribution and estimating the three models with Gamma$(2, 4.1/2)$ infectious period distribution, while in section 8.10, we simulate three dimensional household epidemic with Gamma$(2, 4.1/2)$ infectious period distribution and estimate the three models parameters with exp(4.1) infectious period distribution. We discussed our results in section 8.11.

In section 8.13, we studied the effects of misspecification on the estimates of the three model parameters in the face of misclassification error in the epidemic data with different misclassification probabilities. We simulate four dimensional household epidemic with exp(4.1) infectious period distribution and estimate the three model parameters with Gamma$(2, 4.1/2)$ infectious period distribution. While in section 8.14, we discussed the results.

8.2 Simulating epidemic data with $\exp(4.1)$ and estimating model parameters with $\text{Gamma}(2, 4.1/2)$ infectious period distributions.

We simulate two dimensional model epidemic data with exp(4.1) infectious period distribution and estimated the model parameters with the Gamma$(2, 4.1/2)$ infectious period distribution. Plots of the estimates and tables of mean, standard deviation and root mean square errors are presented.

From figures 8.1 (a)-(d), we see that the estimates are biased and imprecise.

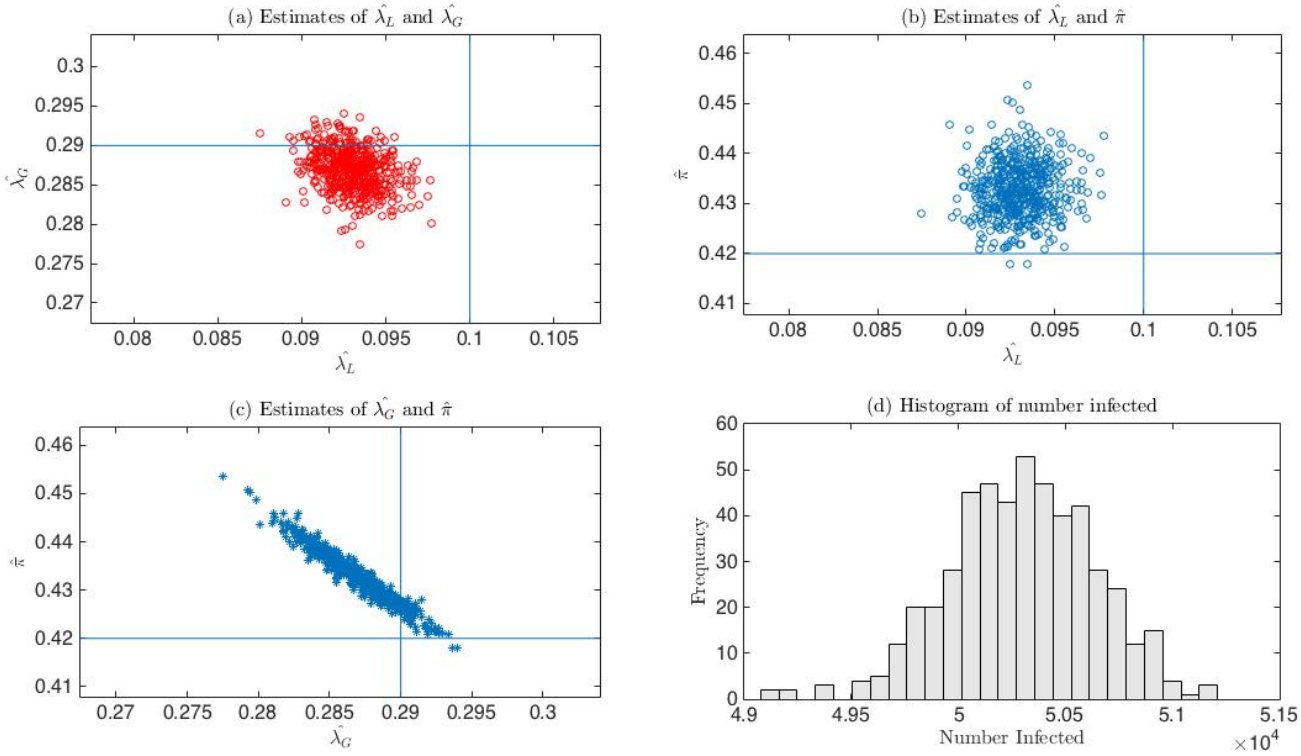

Figure 8.1: Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(1.4) infectious period distribution.

Par. Mean SD, RMSE	Gamma$(2, 4.1/2)$ infectious period distribution				
	λ_L	λ_G	π	z	R_*
Theoretical parameter	0.1	0.29	0.4199	0.7298	2.2166
Mean	0.092993	0.2869	0.43285	0.7119	2.1339
Standard deviation	0.0015132	0.0026017	0.005495	0.0048041	0.019735
Root mean square error	0.0071679	0.0040445	0.014025	0.018455	0.084961

Table 8.1: Table of mean, standard deviation and root mean square error of the estimates when the epidemic data is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions.

8.3 Simulating epidemic data with Gamma$(2, 4.1/2)$ and estimating model parameters with exp(4.1) infectious period distributions.

Here, we estimate the model parameters with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1)$ infectious period distribution.

Plots of the parameter estimates, table of mean, standard and root mean square of the estimates are presented as follows.

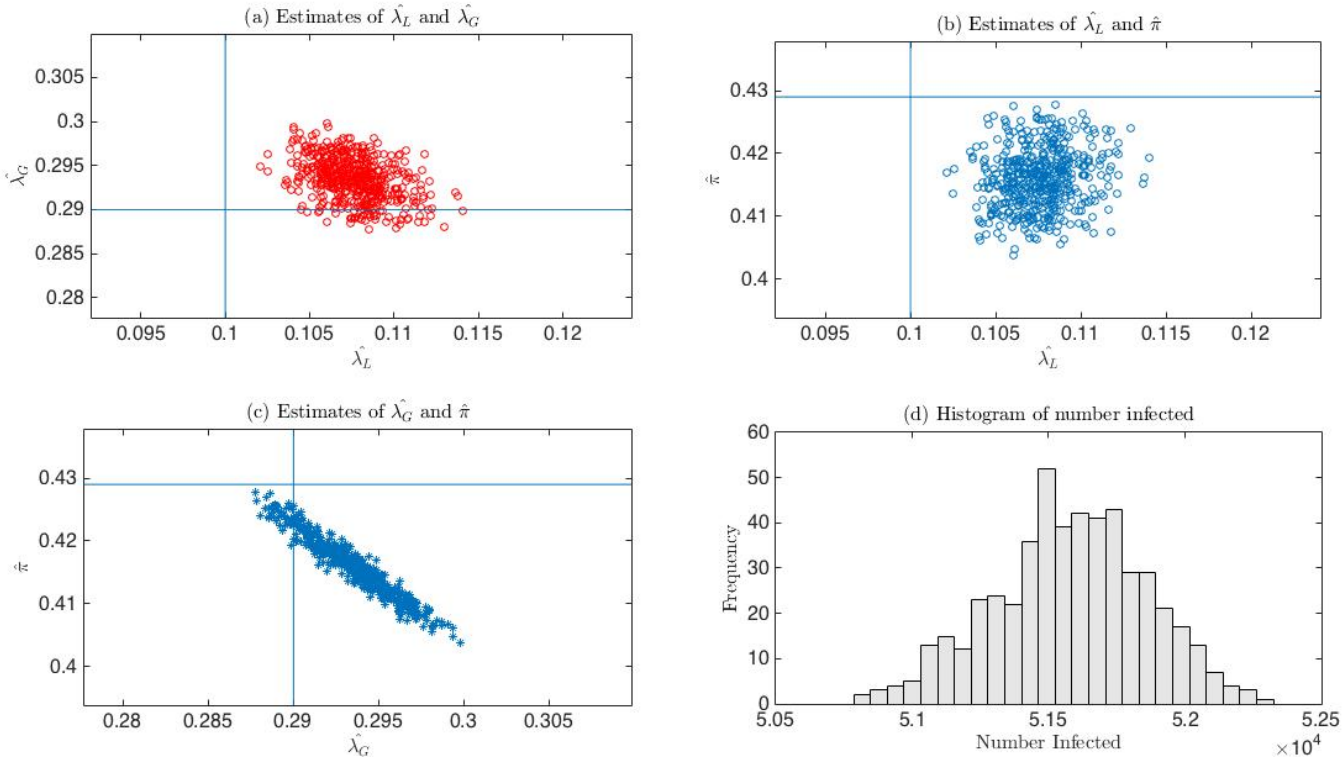

Figure 8.2: Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution.

In figures 8.2 (a)-(d), the estimates are biased and imprecise.

Par. Mean SD, RMSE.	exp(4.1) infectious period distribution				
	λ_L	λ_G	π	z	R_*
Theoretical parameter.	0.1	0.29	0.4291	0.7117	2.1106
Mean	0.10761	0.29351	0.41595	0.72898	2.1878
Standard deviation	0.0019244	0.0023979	0.0047063	0.0039479	0.017285
Root mean square error	0.0078474	0.0042457	0.013884	0.017705	0.079101

Table 8.2: Table of mean, standard deviation and root mean square error of the estimates when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions.

8.4 Discussion and comments.

From figures 8.1 (a)-(d) and figures 8.2 (a)-(d), we see that the estimates from both scenarios are not scattered about their true parameters values as expected but are biased and imprecise.

8.5 Effects of misspecification on the estimates of the three models from two dimensional epidemic data.

We examined the precision of the estimates from the three models in the face of misspecification when the number of infectives and susceptibles are the true number of positives and true number of negatives using pair of theoretical parameters $(\lambda_L, \lambda_G) = (0.0446, 0.1955)$ and $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and large population size to allow global infection in our simulations.

We plot the estimates of the parameters and compute their mean, standard deviation, root mean square error.

Starting with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ we simulate two dimensional household epidemic data with exp(4.1) infectious period distribution and estimate the model parameters with Gamma$(2, 4.1/2)$ infectious period distribution.

Also from figures 8.3 (a)-(e), we see that the estimates from the three models are biased and imprecise with large variability from the four dimensional model. The three and four dimensional models are not significantly better than the two dimensional model.

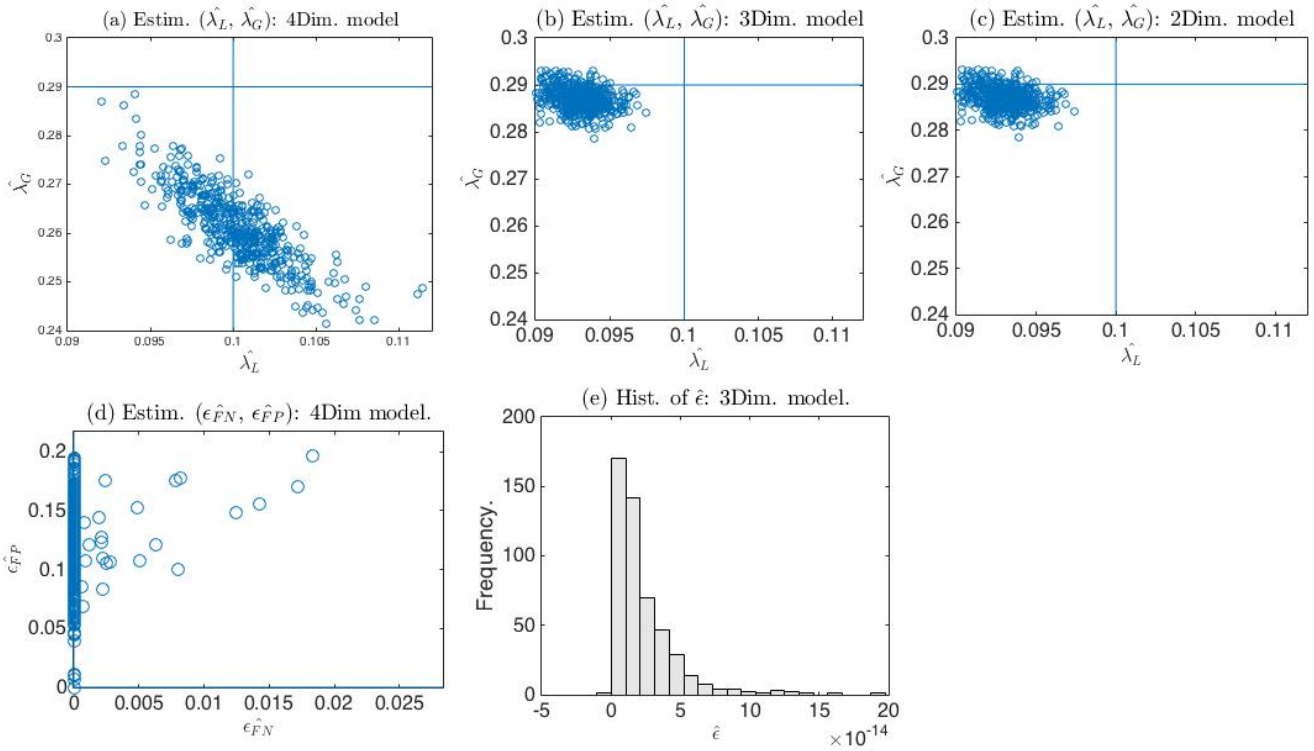

Figure 8.3: Plots of the estimates from the three models when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, the parameters estimated with Gamma$(2, 4.1/2)$ infectious period distribution.

Par. Estim.	Mean of the estimates.			
	2Dim.	3Dim.	4Dim.	Theoret. Param.
$\hat{\lambda_L}$	0.09297	0.09297	0.10036	0.1
$\hat{\lambda_G}$	0.28692	0.28692	0.26082	0.29
$\hat{\pi}$	0.43282	0.43282	0.4873	0.4199
$\hat{z}$	0.7119	0.7119	0.67244	0.7298
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.00024992	N/A
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.11889	N/A
$\hat{\varepsilon}$	N/A	2.28E-14	N/A	N/A
$\hat{R_*}$	2.1339	2.1339	1.9955	2.266

Table 8.3: Table of mean of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution and estimated with Gamma$(2, 4.1/2)$ infectious period distribution.

Par. Estim.	Standard deviation of the estimates.		
	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0014247	0.0014247	0.0029347
$\hat{\lambda_G}$	0.0024752	0.0024752	0.0075469
$\hat{\pi}$	0.0053002	0.0053002	0.016486
$\hat{z}$	0.7119	0.0047022	0.012503
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.0016047
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.0016047
$\hat{\varepsilon}$	N/A	2.36E-14	N/A
$\hat{R_*}$	0.019347	0.019347	0.041717

Table 8.4: Table of standard deviation of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, estimated with Gamma$(2, 4.1/2)$ infectious period distribution.

Par. Estim.	Root mean square error of the estimates.		
	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0071728	0.0071728	0.0029538
$\hat{\lambda_G}$	0.0039488	0.0039488	0.030139
$\hat{\pi}$	0.013967	0.013967	0.069384
$\hat{z}$	0.018502	0.018502	0.058698
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.0016225
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.12282
$\hat{\varepsilon}$	N/A	3.28E-14	N/A
$\hat{R_*}$	0.084906	0.084906	0.22497

Table 8.5: Table of the root mean square error of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and exp(4.1) infectious period distribution, estimated with Gamma$(2, 4.1/2)$ infectious period distribution.

8.6 When the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution.

From figures 8.4 (a)-(e), the estimates of the three models are biased and imprecise with large variability from the four dimensional model.

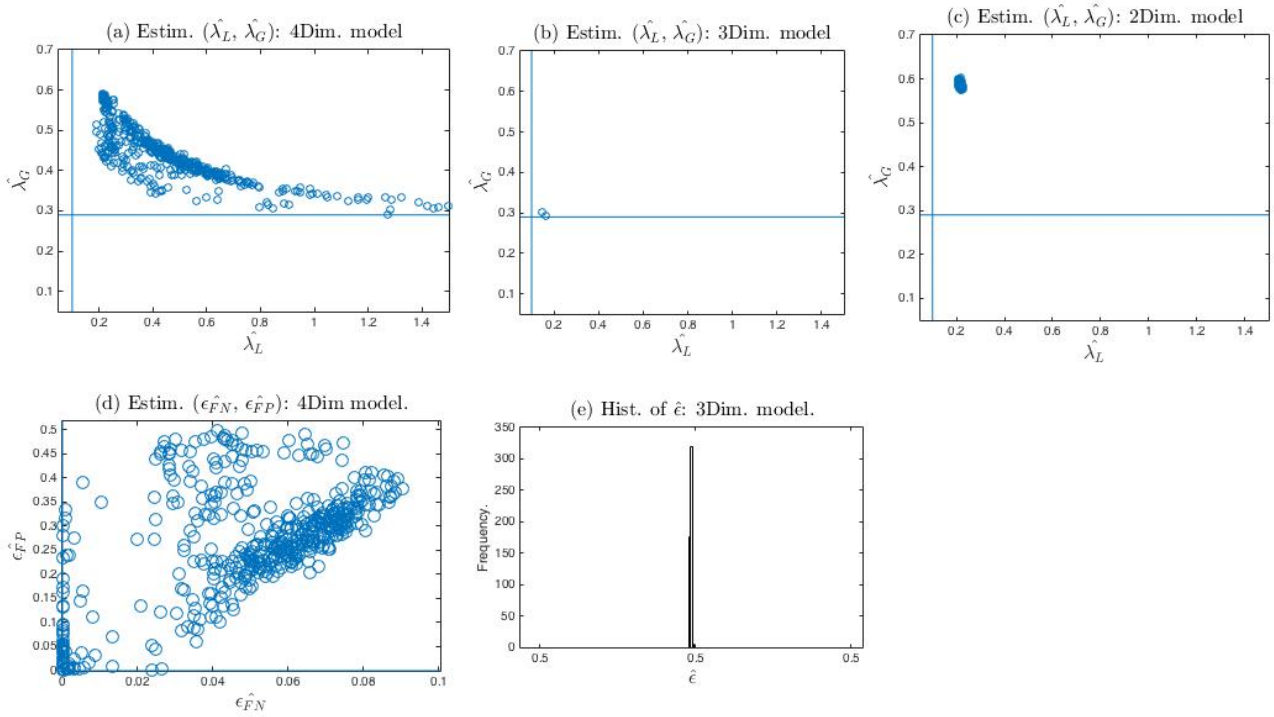

Figure 8.4: Plots of the estimates from the three models when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, the parameters estimated with exp(4.1) infectious period distribution.

Par. Estim.	Mean of the estimates.			
	2Dim.	3Dim.	4Dim.	Theoret. Param.
$\hat{\lambda_L}$	0.21548	0.16046	0.48041	0.1
$\hat{\lambda_G}$	0.58692	0.2954	0.44415	0.29
$\hat{\pi}$	0.41596	0.99995	0.54458	0.4291
$\hat{z}$	0.72908	8.25E-05	0.67212	0.7117
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.048296	N/A
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.25923	N/A
$\hat{\varepsilon}$	N/A	0.5	N/A	N/A
$\hat{R_*}$	2.1883	1.0001	1.9747	2.1106

Table 8.6: Table of mean of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution.

Par. Estim.	Standard deviation of the estimates.		
	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0035179	0.007419	0.32168
$\hat{\lambda_G}$	0.0047777	0.004534	0.072247
$\hat{\pi}$	0.0047436	4.27E-05	0.070432
$\hat{z}$	0.72908	6.83E-05	0.049919
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.025196
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.025196
$\hat{\varepsilon}$	N/A	1.33E-15	N/A
$\hat{R_*}$	0.017345	4.47E-05	0.15808

Table 8.7: Table of standard deviation of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution.

Par. Estim.	Root mean square error of the estimates.		
	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.11553	0.060908	0.49798
$\hat{\lambda_G}$	0.29695	0.0070472	0.17021
$\hat{\pi}$	0.58404	5.17E-05	0.46081
$\hat{z}$	0.72906	8.58E-05	0.67394
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.054461
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.28756
$\hat{\varepsilon}$	N/A	0.5	N/A
$\hat{R_*}$	1.3379	0.14952	1.1352

Table 8.8: Table of root mean square error of the parameter estimates when the epidemic is simulated with theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution, estimated with exp(4.1) infectious period distribution.

8.7 Discussion and comments.

From the plots of the estimates and table of mean, standard deviation and the root mean square error, we see that the estimates from the three models are dispproportionate in values when compared to their true parameter values. Those of the three and four dimensional models are less precise than those of the two dimensional model.

In general, with only misspecification of the model, the two dimensional model is better than the three and four dimensional models on two dimensional epidemic data.

8.8 Misspecification in the face of misclassification.

If the epidemic data is misclassified having the same misclassification probabilities and the model is also misspecified such that the infectious period distribution used in estimation is different from that used in simulating the epidemic data, then how does this affect the precision of the parameters?

We studied this problem using the large population size and theoretical parameters to allow global infection in our simulations. We therefore considered the pair of theoretical parameters, $(\lambda_L, \lambda_G) = (0.1, 0.29)$.

We present plots of the estimates under the two scenarios, in which the epidemic data is

estimated with a different infectious period from that used in simulating the data.

8.9 When the epidemic data is simulated with exp(4.1) and estimated with Gamma(2, 4.1/2) infectious period distributions.

Here we simulate the epidemic data with exp(4.1) infectious period distribution and estimate the model parameters with Gamma(2, 4.1/2) infectious period distribution as follow.

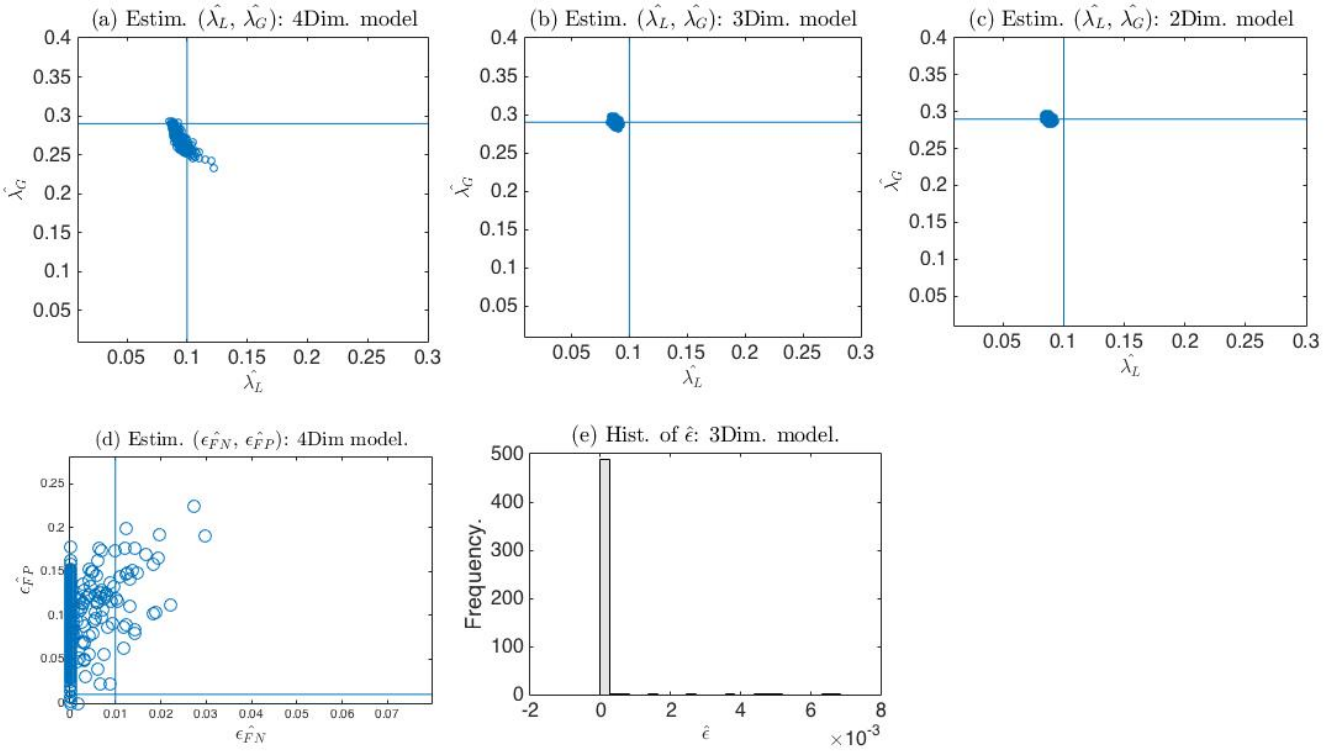

Figure 8.5: Plots of the estimates with Gamma(2, 4.1/2) infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.01$.

In figures 8.5 (a)-(e), the estimates are biased and imprecise with less variability from the two and three dimensional models as shown in figures 8.5 (b) and (c) owing to misspecification. The three dimensional model is better than the two and four dimensional models.

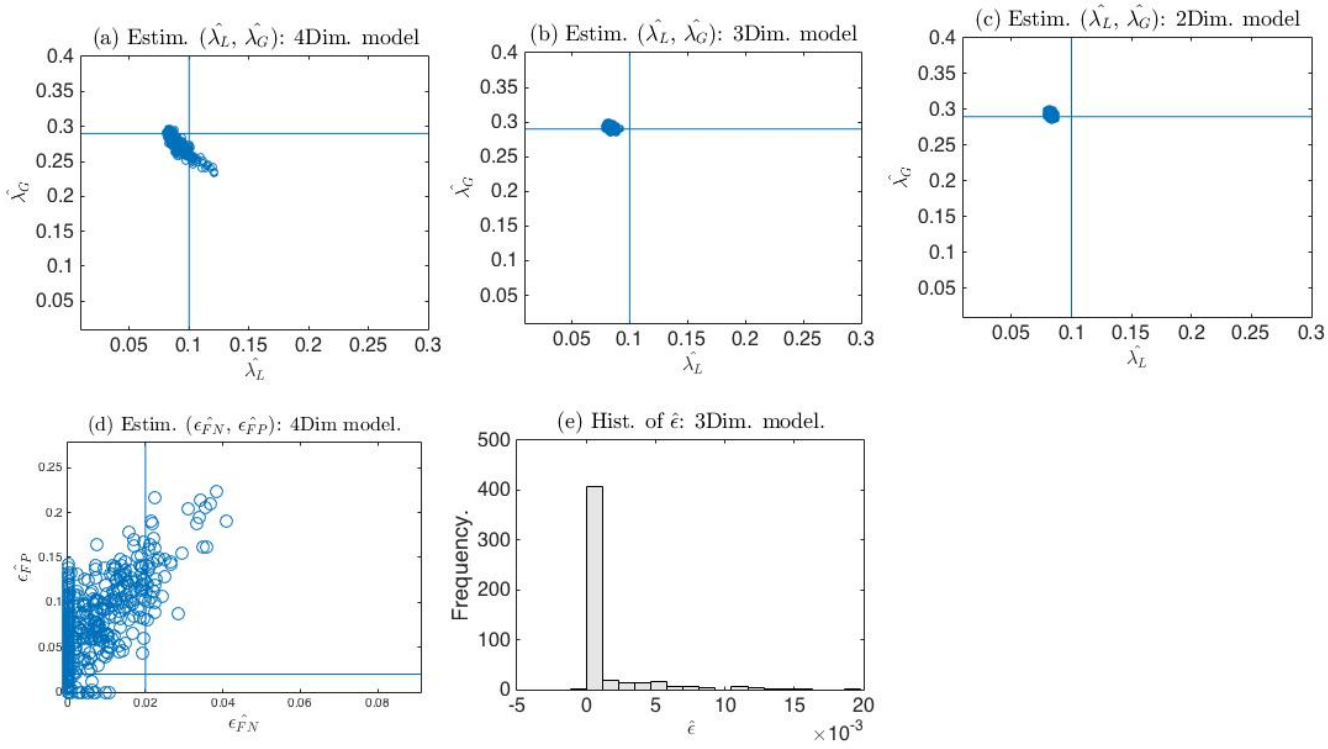

Figure 8.6: Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.02$.

Similar pattern of behaviours in figures 8.5 (a)-(c) can be seen in figures 8.6 (a)-(c). The three dimensional model is better than the two and four dimensional models.

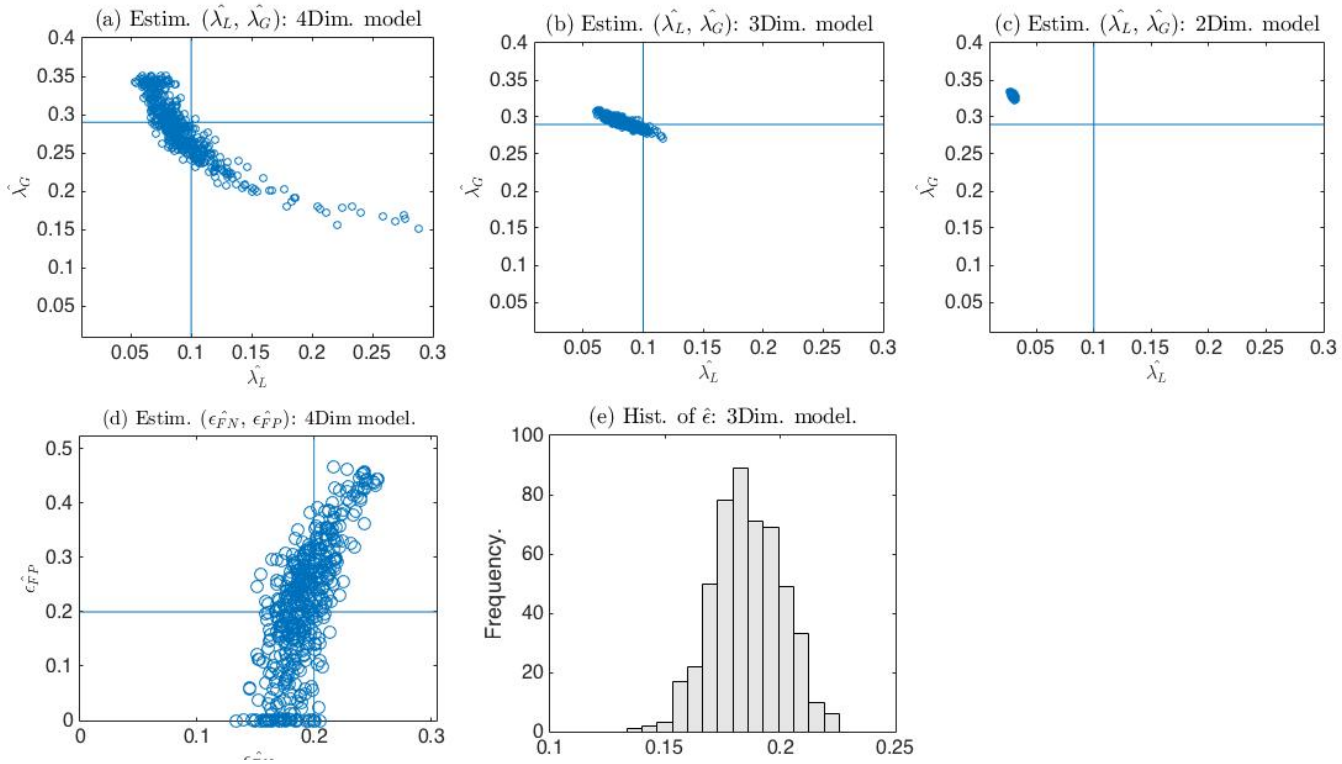

Figure 8.7: Plots of the estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon = 0.2$.

In figures 8.7 (a)-(e), we see large variability of the estimates of the four dimensional model around their true values compared to those of the three dimensional model. While those of the two dimensional model are biased and imprecise. In general the three dimensional model is better than the two and four dimensional models.

	Model									
Par.	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	Theor.
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2	Par.
$\hat{\lambda_L}$	0.087924	0.087965	0.093676	0.083014	0.083485	0.089956	0.029568	0.082831	0.10262	0.1
$\hat{\lambda_G}$	0.28989	0.28987	0.26986	0.29261	0.29234	0.27534	0.32819	0.29288	0.27823	0.29
$\hat{\pi}$	0.43114	0.43115	0.4721	0.43014	0.43025	0.46311	0.43059	0.4303	0.46004	0.4199
$\hat{z}$	0.70788	0.70792	0.67864	0.70323	0.70367	0.6823	0.62622	0.70248	0.68729	0.7298
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.0014387	N/A	N/A	0.0062863	N/A	N/A	0.19136	N/A
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.092036	N/A	N/A	0.077538	N/A	N/A	0.2094	N/A
$\hat{\varepsilon}$	N/A	8.33E-05	N/A	N/A	0.0010209	N/A	N/A	0.18576	N/A	N/A
$\hat{R_*}$	2.1115	2.1117	2.0115	2.0863	2.0887	2.0198	1.7083	2.0823	2.0392	2.2166

Table 8.9: Mean of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the data is simulated with exp(4.1) infectious period distribution.

	Model								
Par.	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.0014051	0.0014386	0.0040913	0.001363	0.0018231	0.0066354	0.00093879	0.0097516	0.065447
$\hat{\lambda_G}$	0.0025843	0.0026028	0.0086708	0.0024284	0.002515	0.010904	0.0020427	0.0061165	0.045795
$\hat{\pi}$	0.005432	0.0054391	0.017883	0.0049548	0.0049653	0.020602	0.0036992	0.0063326	0.084384
$\hat{z}$	0.0047356	0.0047228	0.012985	0.0042594	0.0044268	0.013337	0.0031339	0.010575	0.047823
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.0039525	N/A	N/A	0.008371	N/A	N/A	0.02099
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.036955	N/A	N/A	0.044611	N/A	N/A	0.1174
$\hat{\varepsilon}$	N/A	0.00062226	N/A	N/A	0.0026397	N/A	N/A	0.014962	N/A
$\hat{R_*}$	0.018951	0.018903	0.043454	0.016555	0.017794	0.043505	0.0079267	0.053799	0.16304

Table 8.10: Standard deviation of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the data is simulated with exp(4.1) infectious period distribution.

	Model								
Par.	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.012157	0.012121	0.0075297	0.01704	0.016615	0.012035	0.070438	0.01974	0.065434
$\hat{\lambda_G}$	0.002584	0.0026035	0.0075297	0.0035642	0.0034355	0.012035	0.038244	0.0067555	0.065434
$\hat{\pi}$	0.012492	0.012504	0.055177	0.011382	0.011484	0.047868	0.011316	0.012178	0.093372
$\hat{z}$	0.022415	0.022377	0.052767	0.026896	0.02649	0.049329	0.10362	0.029289	0.063947
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.0094462	N/A	N/A	0.016062	N/A	N/A	0.022678
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.009428	N/A	N/A	0.072779	N/A	N/A	0.11766
$\hat{\varepsilon}$	N/A	0.0099362	N/A	N/A	0.019161	N/A	N/A	0.020643	N/A
$\hat{R_*}$	0.10678	0.10658	0.20961	0.13136	0.12914	0.20151	0.50835	0.14461	0.24079

Table 8.11: Root mean square error of the parameter estimates with Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution.

8.10 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions.

We examined the properties of the estimates under these scenarios, presented their plots and tables of mean standard deviation and root mean square error.

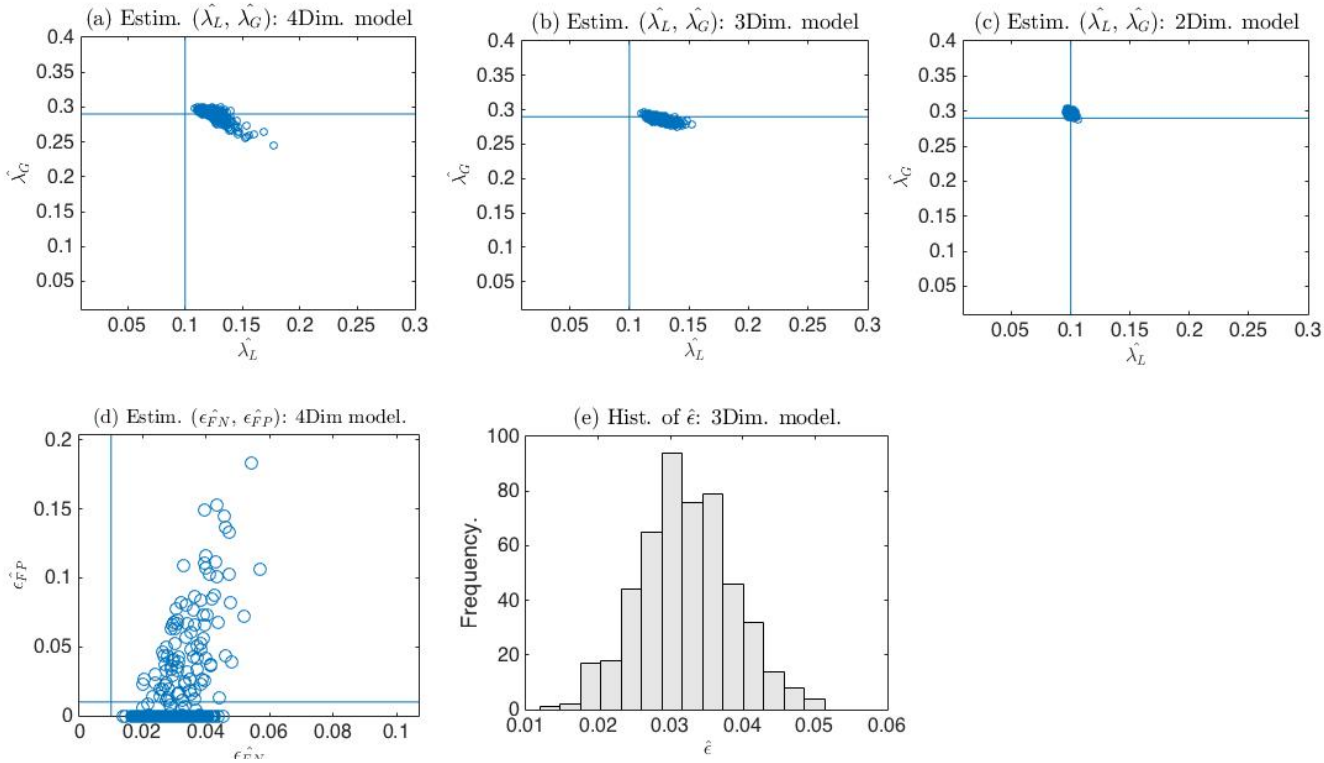

Figure 8.8: Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.01$.

In figure 8.12 (c), the estimates of the two dimensional model are more precise with less variability than the three and four dimensional models in figures 8.12 (a) and (b) owing to misspecification.

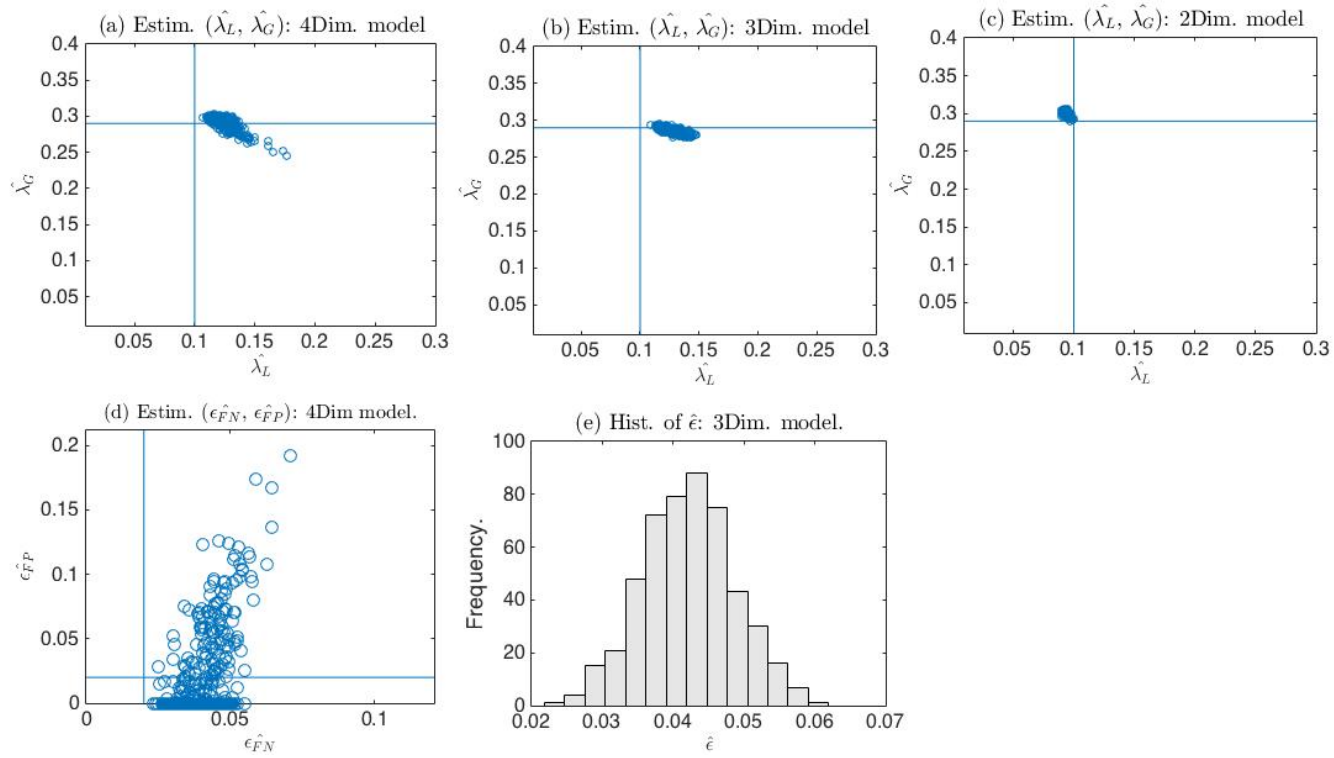

Figure 8.9: Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.02$.

Similar behaviours in figures 8.12 (a)-(e) can be seen in figures 8.13 (a)-(e).

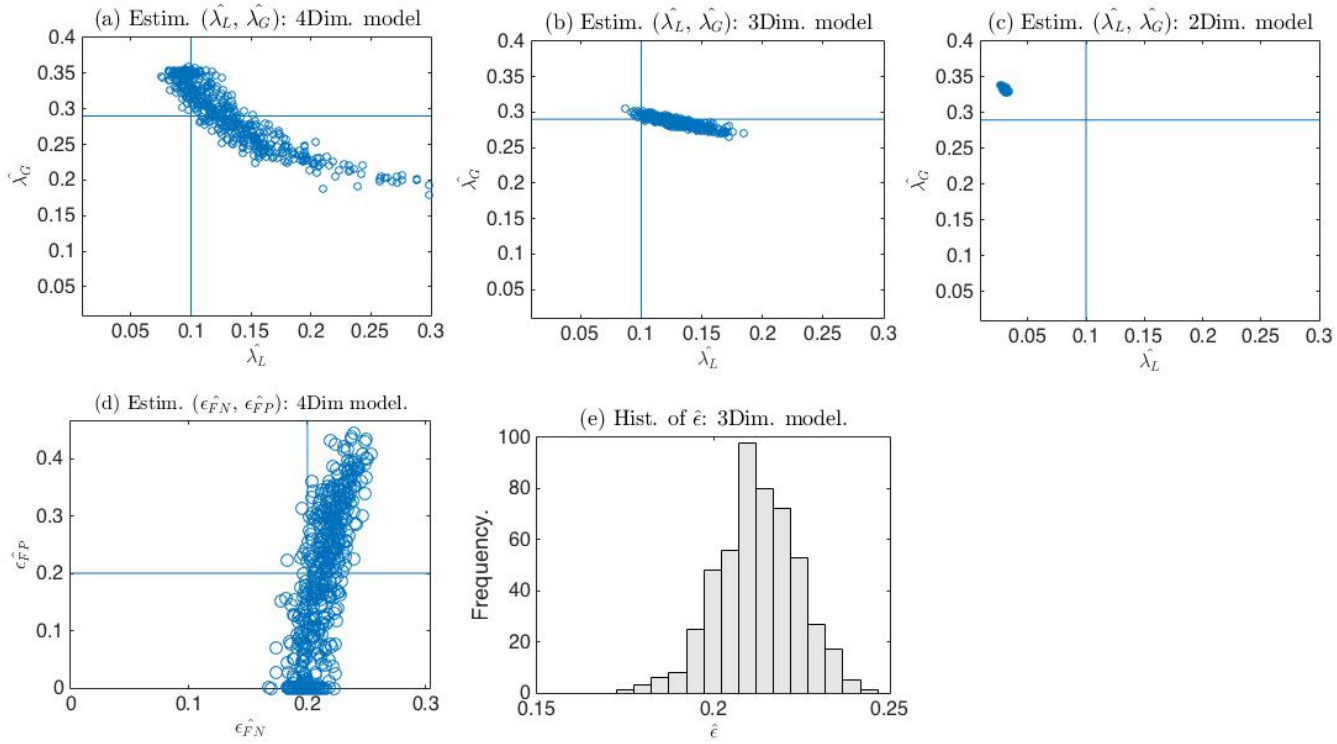

Figure 8.10: Plots of the estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon = 0.2$.

In figures 8.10 (a)-(e), the estimates of the three and four dimensional models are centered at their true values with more variability for the four dimensional model than those of the three dimensional model. While those of the two dimensional model are imprecise and biased.

In general the three dimensional model is better than the two and four dimensional models

	Model									
Par.	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	Theor.
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2	Par.
$\hat{\lambda_L}$	0.10058	0.1274	0.12439	0.094141	0.12782	0.12418	0.030468	0.12778	0.14371	0.1
$\hat{\lambda_G}$	0.29647	0.28606	0.29078	0.29931	0.28595	0.29154	0.333	0.28622	0.28769	0.29
$\hat{\pi}$	0.41465	0.41957	0.41156	0.41355	0.41959	0.41022	0.41912	0.41963	0.42183	0.4291
$\hat{z}$	0.72426	0.74059	0.74488	0.71956	0.74085	0.74575	0.63694	0.74022	0.74092	0.7117
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.030553	N/A	N/A	0.040032	N/A	N/A	0.21275	N/A
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.010832	N/A	N/A	0.01722	N/A	N/A	0.18887	N/A
$\hat{\varepsilon}$	N/A	0.032266	N/A	N/A	0.042179	N/A	N/A	0.21254	N/A	N/A
$\hat{R_*}$	2.1617	2.2509	2.2698	2.1356	2.2522	2.2738	1.73	2.2486	2.2593	2.1106

Table 8.12: Mean of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution.

	Model								
Par.	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.0017218	0.0064309	0.0076548	0.0016489	0.0068501	0.0087742	0.00098638	0.016316	0.070344
$\hat{\lambda_G}$	0.0024134	0.0033416	0.0069713	0.0023772	0.0033419	0.0082493	0.0019164	0.0061721	0.04847
$\hat{\pi}$	0.0047786	0.0052086	0.012119	0.004571	0.0050424	0.014044	0.0033002	0.0065439	0.083907
$\hat{z}$	0.0040338	0.0051911	0.007914	0.0037568	0.0053171	0.008631	0.0027658	0.010044	0.044281
$\varepsilon\hat{_{FN}}$	N/A	N/A	0.0064402	N/A	N/A	0.0069957	N/A	N/A	0.015604
$\varepsilon\hat{_{FP}}$	N/A	N/A	0.027343	N/A	N/A	0.031997	N/A	N/A	0.12798
$\hat{\varepsilon}$	N/A	0.0063887	N/A	N/A	0.0065142	N/A	N/A	0.011181	N/A
$\hat{R_*}$	0.017093	0.025695	0.037354	0.01539	0.026481	0.040686	0.0074831	0.052057	0.19753

Table 8.13: Standard deviation of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution.

Par.	Model								
	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim	2Dim	3Dim	4Dim
ε	0.01	0.01	0.01	0.02	0.02	0.02	0.2	0.2	0.2
$\hat{\lambda_L}$	0.0018158	0.028146	0.025563	0.0060862	0.028645	0.025722	0.069539	0.032211	0.082761
$\hat{\lambda_G}$	0.006906	0.0051639	0.025563	0.009607	0.0052508	0.025722	0.043042	0.0072334	0.082761
$\hat{\pi}$	0.015189	0.010833	0.021286	0.016179	0.010735	0.023495	0.01048	0.011482	0.084135
$\hat{z}$	0.013232	0.0294	0.034151	0.008749	0.02967	0.035171	0.074762	0.030281	0.053042
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.021537	N/A	N/A	0.021216	N/A	N/A	0.020137
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.027328	N/A	N/A	0.032086	N/A	N/A	0.12833
$\hat{\varepsilon}$	N/A	0.023163	N/A	N/A	0.023114	N/A	N/A	0.016796	N/A
$\hat{R_*}$	0.053869	0.14262	0.16354	0.02936	0.14404	0.16818	0.38065	0.14746	0.24705

Table 8.14: Root mean square error of the parameter estimates with exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution.

8.11 Conclusion and comments.

With misclassification error in the data and misspecification, the estimates of the three dimensional model are biased with less variability around their true values than those of the four dimensional model. Those of the two dimensional model are biased and imprecise.

In general, the three dimensional model is better than the two and four dimensional models on three dimensional epidemic data with model misspecification.

8.12 Misspecification in the face of different misclassification Probabilities.

Here, we studied the effect of misspecification on the estimate of the model parameters, when the epidemic data is misclassified with different misclassification probabilities, such that the infectious period distribution used in estimation is different from that used in simulating the epidemic data.

We examined this problem by simulating epidemic data with the pair of theoretical parameters $(\lambda_L, \lambda_G) = (0.1, 0.29)$ and Gamma$(2, 4.1/2)$ infectious period distribution and estimating the models with exp(4.1) infectious period distributions and vice versa.

8.13 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with exp(4.1) and estimated with Gamma(2, 4.1/2) infectious period distributions.

We simulate epidemic data with exp(4.1) infectious period distribution and estimate the model with Gamma(2, 4.1/2) infectious period distribution. We present plots of the estimates and table of the mean, standard deviation and root mean square error.

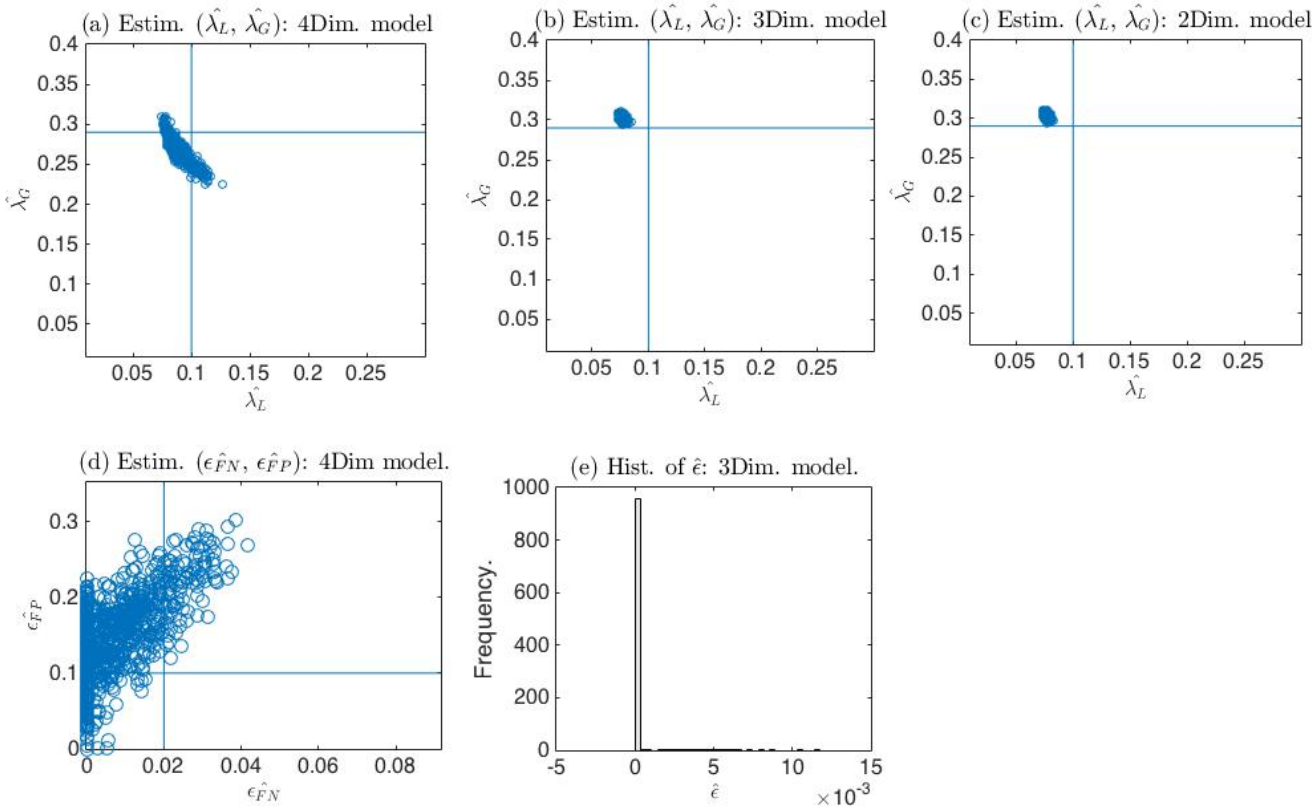

Figure 8.11: Plots of the estimates using Gamma(2, 4.1/2) infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.1$.

In figures 8.11 (a)-(c), the estimates of the three models are biased and imprecise.

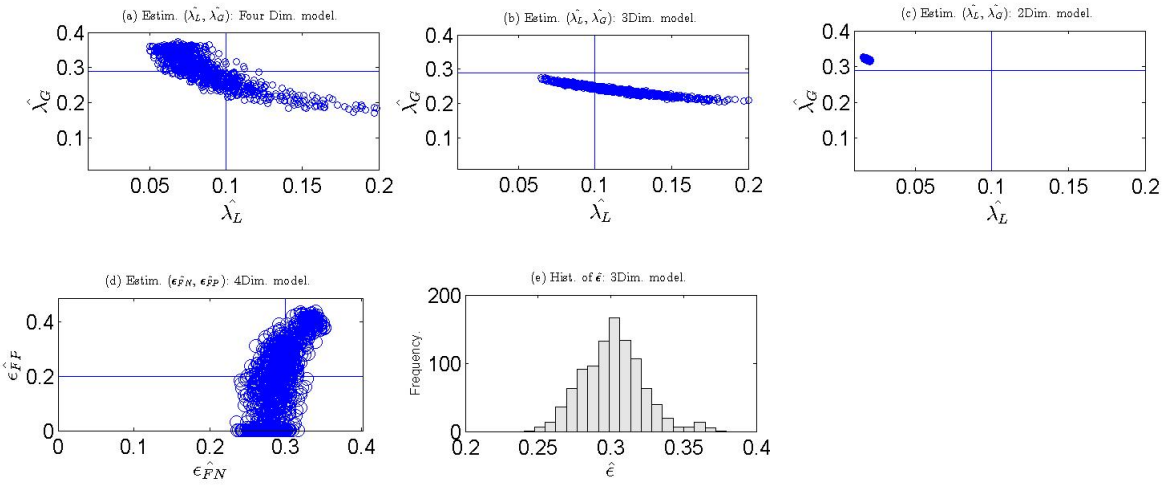

Figure 8.12: Plots of the estimates using Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.3$, $\varepsilon_{FF} = 0.2$.

In figures 8.12 (a) and (b), we see large variability of the estimates of the three and four dimensional models around their true values. While those of the two dimensional model are biased and imprecise. The three and four dimensional models are better than the two dimensional model.

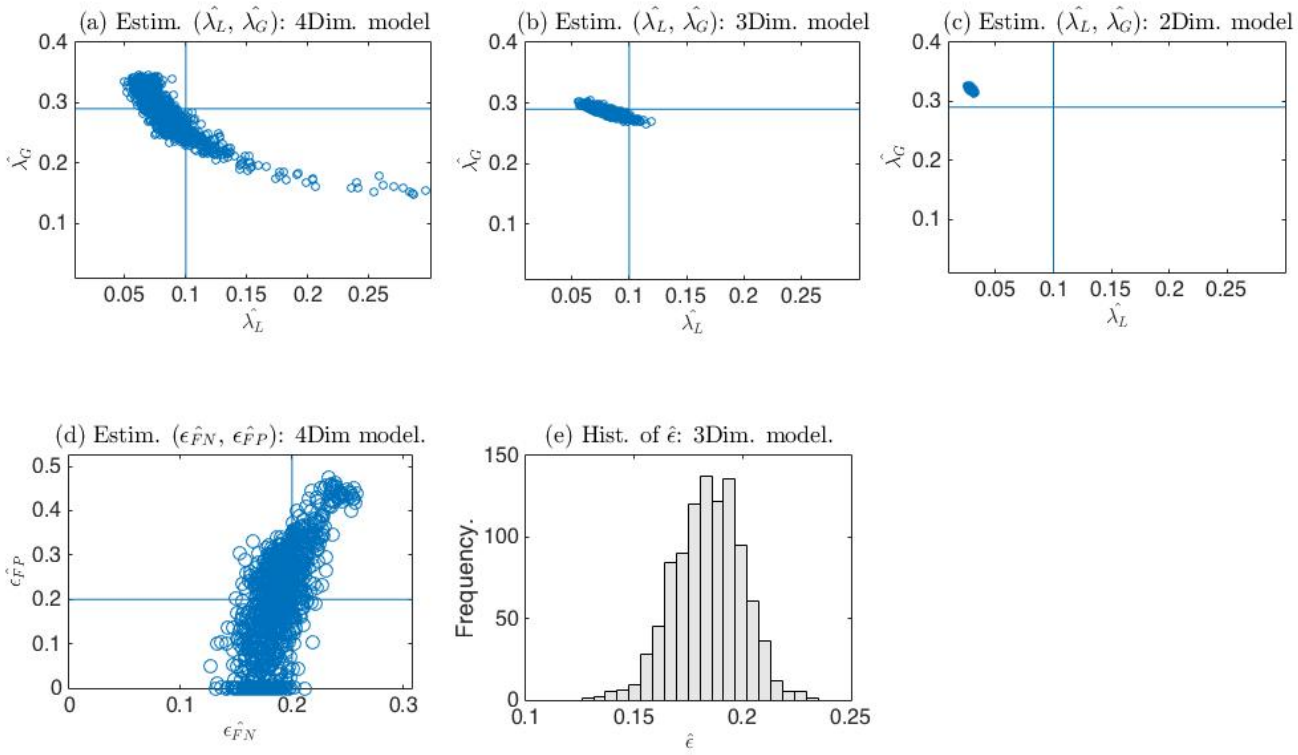

Figure 8.13: Plots of the estimates using Gamma$(2, 4.1/2)$ infectious period distribution when the epidemic data is simulated with exp(4.1) infectious period distribution and $\varepsilon_{FN} = 0.2$, $\varepsilon_{FF} = 0.2$.

In figures 8.13 (a) and (b), the scatter plots of the estimates of the four and three dimensional models are centered around their true values but with more variability from those of the four dimensional model. While those of the two dimensional model are biased and imprecise. Given this scenario, the three dimensional model is significantly better than the two and four dimensional models as theoretically expected.

Par. Estim.	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			Theo.
	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	Param.
$\hat{\lambda_L}$	0.077163	0.077248	0.087328	0.018492	0.12408	0.11088	0.1
$\hat{\lambda_G}$	0.30254	0.3025	0.27016	0.32022	0.23888	0.28331	0.29
$\hat{\pi}$	0.39733	0.39734	0.46066	0.48234	0.53371	0.45364	0.4199
$\hat{z}$	0.7264	0.7265	0.68375	0.55535	0.64279	0.69246	0.7298
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.0068771	N/A	N/A	0.29176	N/A
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.14662	N/A	N/A	0.18414	N/A
$\hat{\varepsilon}$	N/A	0.00019064	N/A	N/A	0.30141	N/A	N/A
$\hat{R_*}$	2.1701	2.1707	2.0246	1.5303	1.8966	2.0626	2.2166

Table 8.15: Table of mean of the parameter estimates when the epidemic is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions.

Par. Estim.	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$		
	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0013641	0.0014384	0.0075362	0.00083274	0.098363	0.085154
$\hat{\lambda_G}$	0.0025697	0.0025712	0.0025712	0.0018798	0.017461	0.06154
$\hat{\pi}$	0.0050036	0.005004	0.026227	0.003523	0.019542	0.11398
$\hat{z}$	0.0042381	0.0042876	0.017032	0.0029415	0.017353	0.06668
$\hat{\varepsilon_{FN}}$	N/A	N/A	0.0089027	N/A	N/A	0.022678
$\hat{\varepsilon_{FP}}$	N/A	N/A	0.0089027	N/A	N/A	0.022678
$\hat{\varepsilon}$	N/A	0.0010166	N/A	N/A	0.02093	N/A
$\hat{R_*}$	0.018388	0.018763	0.055725	0.0054852	0.051097	0.23005

Table 8.16: Table of standard deviation the parameter estimates when the epidemic is simulated with exp(4.1) and estimated withGamma$(2, 4.1/2)$ infectious period distributions.

8.13.1 Plots of the estimates and table of mean, standard deviation, root mean square error when the epidemic data is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions.

We simulate epidemic data with Gamma$(2, 4.1/2)$ infectious period distribution and estimate the model with exp(4.1) infectious period distribution. We then present plots of the estimates and table of mean, standard deviation and root mean square error.

In figures 8.14 (a)-(c), the scatter plots of the estimates of λ_L and λ_G from the three and four dimensional models are close to their true values with more variability from those of the

Par. Estim.	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$		
	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.022877	0.022797	0.014742	0.081513	0.10122	0.085804
$\hat{\lambda_G}$	0.012802	0.01276	0.023996	0.030277	0.054021	0.085804
$\hat{\pi}$	0.0077887	0.0077836	0.063068	0.062538	0.11547	0.11881
$\hat{z}$	0.019558	0.019481	0.064056	0.17446	0.088712	0.076388
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.015855	N/A	N/A	0.024118
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.070549	N/A	N/A	0.13442
$\hat{\varepsilon}$	N/A	0.059818	N/A	N/A	0.0555	N/A
$\hat{R_*}$	0.1225	0.12201	0.27237	0.68633	0.32402	0.27674

Table 8.17: Table of root mean square error of the parameter estimates when the epidemic is simulated with exp(4.1) and estimated with Gamma$(2, 4.1/2)$ infectious period distributions.

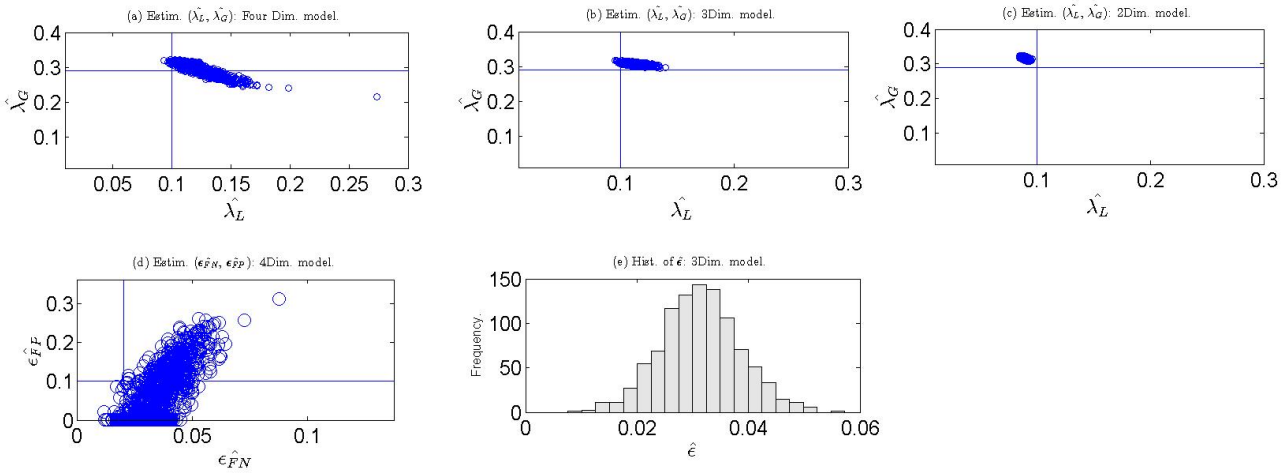

Figure 8.14: Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.1$.

four dimensional model. While those of the two dimensional model are biased.

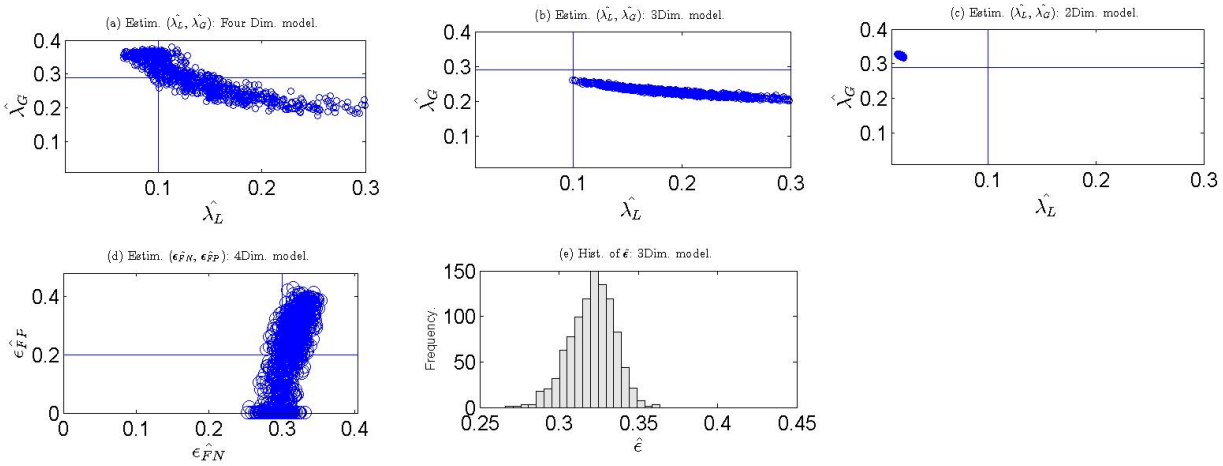

Figure 8.15: Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$.

In figures 8.15, (a)-(e), similar behaviours in figures 8.14 (a)-(c) are shown with less variability.

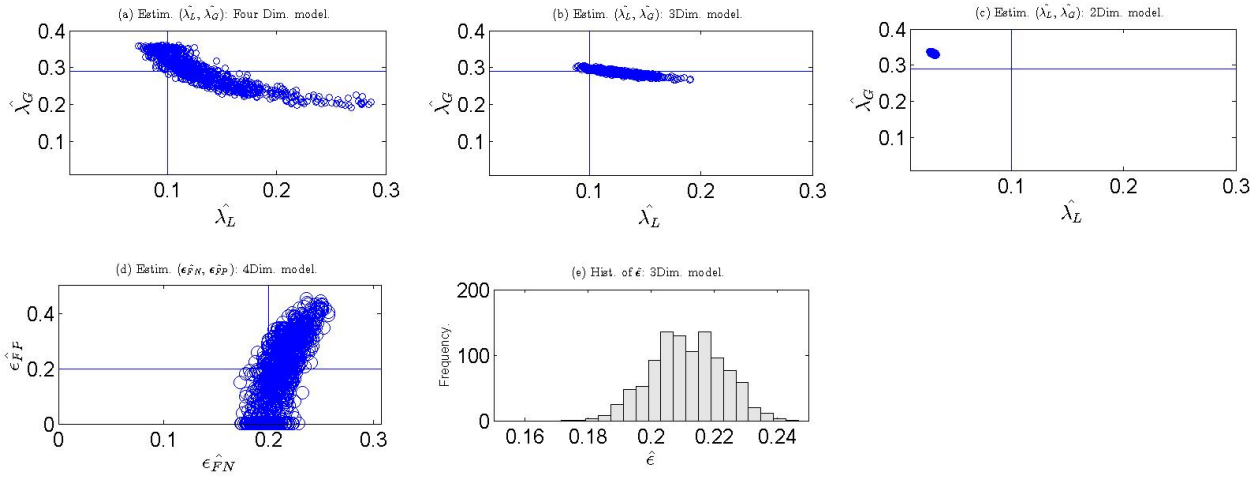

Figure 8.16: Plots of the estimates using exp(4.1) infectious period distribution when the epidemic data is simulated with Gamma$(2, 4.1/2)$ infectious period distribution and $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.2$.

Also, similar behaviours in figures 8.13 (a)-(d) are repeated in figures 8.16 (a)-(d)

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$			Theo.
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	Param.
$\hat{\lambda_L}$	0.0899	0.11325	0.12146	0.018857	0.20678	0.17525	0.030476	0.12636	0.14328	0.1
$\hat{\lambda_G}$	0.31604	0.3072	0.29637	0.32328	0.22549	0.28042	0.33295	0.28666	0.28834	0.29
$\hat{\pi}$	0.38268	0.38487	0.403	0.47353	0.53291	0.43831	0.41922	0.41955	0.42131	0.4291
$\hat{z}$	0.74134	0.75816	0.74916	0.564	0.68191	0.7314	0.63686	0.73928	0.74019	0.7117
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.035793	N/A	N/A	0.31001	N/A	N/A	0.21176	N/A
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.073987	N/A	N/A	0.19614	N/A	N/A	0.18777	N/A
$\hat{\varepsilon}$	N/A	0.031352	N/A	N/A	0.32081	N/A	N/A	0.21156	N/A	N/A
$\hat{R_*}$	2.2213	2.3271	2.2885	1.5441	2.0035	2.221	1.7299	2.2436	2.2554	2.1106

Table 8.18: Table of mean of the parameter estimates when the epidemic is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions..

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$		
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.0016924	0.0066428	0.01531	0.00088043	0.10474	0.33149	0.0010178	0.016721	0.079516
$\hat{\lambda_G}$	0.0025711	0.0033824	0.017484	0.0018071	0.013616	0.060984	0.0020738	0.0064708	0.048072
$\hat{\pi}$	0.0044885	0.0047762	0.029084	0.0031539	0.016093	0.10675	0.0035174	0.0069734	0.083025
$\hat{z}$	0.0034881	0.0055627	0.015519	0.0025597	0.014854	0.057308	0.0027811	0.010356	0.043984
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.0096162	N/A	N/A	0.017639	N/A	N/A	0.01552
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.0096162	N/A	N/A	0.013944	N/A	N/A	0.01552
$\hat{\varepsilon}$	N/A	0.0070962	N/A	N/A	0.32081	N/A	N/A	0.19768	N/A
$\hat{R_*}$	0.015789	0.032068	0.070022	0.0051863	0.042476	0.24985	0.0071748	0.05349	0.19768

Table 8.19: Table of standard deviation the parameter estimates when the epidemic is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions.

	$\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1.$			$\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2.$			$\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2.$		
Par. Estim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.
$\hat{\lambda_L}$	0.01024	0.014822	0.026358	0.081148	0.14954	0.33976	0.069531	0.031216	0.090497
$\hat{\lambda_G}$	0.026164	0.01753	0.018601	0.033325	0.065926	0.061702	0.043003	0.00728	0.048076
$\hat{\pi}$	0.046586	0.044434	0.01860	0.044594	0.1051	0.1071	0.010441	0.011785	0.083343
$\hat{z}$	0.029886	0.046832	0.04058	0.14768	0.033251	0.060585	0.07485	0.029491	0.052406
$\varepsilon_{\hat{FN}}$	N/A	N/A	0.018487	N/A	N/A	0.0202759	N/A	N/A	0.019465
$\varepsilon_{\hat{FP}}$	N/A	N/A	0.073775	N/A	N/A	0.13298	N/A	N/A	0.12738
$\hat{\varepsilon}$	N/A	0.029513	N/A	N/A	0.072166	N/A	N/A	0.016198	N/A
$\hat{R_*}$	0.11175	0.21884	0.19116	0.56651	0.11525	0.27301	0.3808	0.14333	0.24492

Table 8.20: Table of root mean square error of the parameter estimates when the epidemic is simulated with Gamma$(2, 4.1/2)$ and estimated with exp(4.1) infectious period distributions.

8.14 Conclusion and comments.

From the scatter plots (a)-(e) in figures 8.11 , 8.12 and 8.13 and tables 8.15, 8.16 and 8.17, we see that estimates of the four dimensional model are more precise than those of the two

and three dimensional models in the face of misspecification when the epidemic data is four dimensional data. Also, we see in figures 8.14-8.16, (a)-(e), that with misspecification the three and four dimensional models are better than the two dimensional models, when the data are both misclassified and misspecified.

Chapter 9

Summary, Conclusion and Extensions.

9.1 Introduction.

In this chapter, we summarised the work done and discussed the results. We also provided inferential procedures for analysing the stochastic SIR household model when the final data epidemic data is misclassified and highlighted aspects that may require further extension. This chapter is organised as follows.

In section 9.2, we summarised the work done, while in section 9.3 we discussed the results. In section 9.4, we examined some of the extensions of the stochastic household epidemic model of [9] and explored the need to adjust our inference in face of misclassification error in the final size epidemic data.

In section 9.5, we discussed the conclusion from the results of our studies, in section 9.6, we discussed the limitations of our studies and finally in section 9.7, we outlined the procedures of analysing the stochastic SIR household epidemic when the final size epidemic data is misclassified.

9.2 Summary of Work.

In chapter 2, we studied [9], the stochastic SIR household epidemic model, its household structure, mixing assumptions, branching process approximation of epidemic in the early stages, the threshold theorem of epidemic, the mean final size of the epidemic and the final

size probabilities.

We examined other extension by [1], which assumed independence of epidemics in each household contrary to [9] assumption of dependency of epidemic between households. Using this assumption in [1], we developed a maximum likelihood estimation algorithm for the estimation of the parameters. We constructed the approximate likelihood function of the parameters and developed Matlab programs to estimate the parameters.

Using these procedures in our simulation studies, we examined the threshold behaviour of the model and found that large epidemics only occur when $R_* > 1$, in accordance with [11] and given R_*, the precise values of λ_L and λ_G have little effect on either the number of people infected or the probability of a large epidemic occurring.

In chapter 3, we studied the theoretical properties of some of the functions of the model and their behaviours near the boundaries of the local infection rate. We see that, without local contacts in the households, everybody avoids local infection and the final size of the epidemic is only the initial infective in the household, while with very large local infection rate, everybody in the household is infected.

The threshold parameter reduces to $R_0 = \lambda_G E(T_I)$, whenever $\lambda_L \to 0$. This is the threshold parameter for a population in which every household has one member, the so called general stochastic epidemic model.

We discussed the distribution of the number of individuals infected in the households, its mean, the proportion of the initial susceptibles ultimately infected z, its governing equation for the single population deterministic SIR epidemic model and its behaviours near the boundaries of the local infection rate.

In chapter 4, we fitted the data from the two dimensional model to the two dimensional final size epidemic model using the modified version of the simhouses simulation package which is embedded with Matlab codes, which are based on the maximum likelihood algorithm of [1] and employed the Nelder-Mead fminsearch numerical optimisation. The modified version of the simhouses simulation package computes the estimate of the parameters of the model using [24] method of obtaining the initial values for the estimates. It also computes the means, standard deviation, mean square error, and root mean square error of the estimates.

Comparison of our estimates with those of [1,9], showed that the estimates are the same to the numerical accuracy used and hence our Matlab programs that are used to estimate them are working well. Also, we explored the choice of the minimum epidemic size in simulations with small and large population sizes, 1414 and 70700 respectively and also examined the effects of an overly large minimum epidemic size.

The estimates of the parameters of the model are further explored for a range of values of z with large population size to provide clarity on their properties.

In chapter 5, we developed the theoretical framework leading to misclassification of epidemic data, where the misclassification probabilities are assumed different from each other

Here, the probabilities of making precise observation of an infective when it is true and a susceptible when it is true have the form, $1-\varepsilon_{FN}$ and $1-\varepsilon_{FP}$, respectively.

The distribution of observing j infectives correctly and incorrectly and that of susceptibles correctly and incorrectly are shown to be Binomial distributed in section 5.2. The distribution of the number of infectives and those of the susceptibles observed are given in equations (5.2.2) and (5.2.3) respectively.

The expressions for $P_{i,j}(n)$ and hence $q_{n,i}$ can accommodate cases in which the misclassification probabilities are the same which we have referred to as the three dimensional model. We examined the precision of the estimates under this scenario, using simulation studies and compared them with those of the two and four dimensional models on the final size epidemic data.

In chapter 6, we analysed further the properties of these models, using the Pearson chi-square goodness of fit and the Kolmogorov-Smirnov goodness of fit tests. We fitted these models to the final size epidemic data and plotted the density histograms of the Pearson chi-square. The density histograms are superimposed with theoretical chi-square distribution function. Also, we plotted the empirical cumulative distribution function and the hypothesized chi-squared distribution function. We computed the mean, variance of the Pearson chi-square statistics for the three models.

We explored the parameter estimates of the models along the diagonal of the misclassification probabilities region $\{(\varepsilon_{FN},\varepsilon_{FP}) : \varepsilon_{FN} = 0.2-\varepsilon_{FP}, \varepsilon_{FP} \in [0,0.2]\}$ and along the vertical

axes by holding ε_{FP} constant while varying $\varepsilon_{FN} \in [0, 0.2]$. We then computed their corresponding Pearson chi-square statistics, their mean and variance and plotted these statistics for varying misclassification probabilities in $[0, 0.2]$, and theoretical parameters corresponding to $z = 0.2144$, 0.7298.

We computed the proportion of the simulations rejected from the Pearson chi-square goodness of fit test and explored them for the three and four dimensional final size epidemic data and theoretical parameters corresponding to $z = 0.2144$ and $z = 0.7298$ respectively for misclassification probabilities in $[0, 0.2)$.

In summary, our studies show that the density histograms of Pearson chi-square statistics from the models sufficiently approximate the theoretical chi-square distribution for the three models on the two dimensional final size epidemic data. Since less complex model fits well to the final size epidemic data and in line with the principle of parsimony, the two dimensional model is the preferred model fit to the final size data.

On the three dimensional final size epidemic data, we see that the Pearson chi-square statistics from the two dimensional model failed to approximate their theoretical counterparts when the misclassification becomes large including the mean and variance of the Pearson chi-square statistics in contrast to the Pearson chi-square statistics from the three and four dimensional models, which remain consistently stable under these scenarios. The plot of the cumulative distribution function provided more clarity on these behaviours.

The Pearson chi-square statistics from the three and four dimensional models sufficiently approximate the theoretical chi-square distribution. Plots of the cumulative distribution function of the chi-square goodness of fit statistics and those of their theoretical chi-square distributions from the models are presented. Hence the three and four dimensional models sufficiently fit the three dimensional final size epidemic data while the two dimensional model failed to fit especially when the misclassification probability is not close to 0.

On the data from the four dimensional epidemic model, the Pearson chi-square statistics from the two and three dimensional models failed to approximate the theoretical chi-square distribution, when the misclassification probabilities are large and far apart from each other as shown in figures 6.6 and 6.7 respectively. We see significantly large values of the mean and

variance in contrast to their theoretical values. Under this scenario, the preferred model fit to the four dimensional final size epidemic data is the four dimensional model.

In chapter 7, we continued our studies of the properties of the three models using simulation studies by employing the chi-square difference statistic and the Kolmogorov-Smirnov goodness of fit tests to check the adequacy of the chi-square approximations of the three model. Also, we studied the mean and variance of the chi-square difference statistic from the three models including the proportion of the simulations rejected from the difference chi-square test. This is done by exploring the estimates of the three models, for a range of misclassification probabilities in the permissible region $[0, 0.5)$.

We see that on the two dimensional final size data, the density histograms of the chi-square difference statistic and the cumulative distribution function of the chi-square difference statistics, $D_{2,3}$, $D_{2,4}$, and $D_{3,4}$ approximate the theoretical chi-square distribution χ^2_1, χ^2_2 and χ^2_1 respectively. These behaviours are in line with the discussion in chapter 6. On the three dimensional final size epidemic data, we found that only $D_{3,4}$ is approximately χ^2_1 in the face of large misclassification probabilities in its permissible region, $[0, 0.5)$, while $D_{2,3} \gg \chi^2_1$ and $D_{2,4} \gg \chi^2_2$. This means that only the three and four dimensional models are sufficient on the three dimensional final size epidemic data for misclassification probabilities in the permissible region, $[0, 0.5)$.

These results are consistent and in agreement with those of our previous studies in chapter 6, in which the three and four dimensional models are found to be sufficient on the three dimensional final size epidemic data if the misclassification probability is not close to 0.

On the four dimensional final size data, we see that if the misclassification probabilities are far apart from each other, then $D_{2,3} \gg \chi^2_1$, $D_{3,4} \gg \chi^2_1$ and $D_{2,4} \gg \chi^2_2$. This means that the four dimensional model fits data from the four dimensional final size epidemic model significantly better than the two and three dimensional models.

The four dimensional model outperforms the two and three dimensional models on the final size epidemic data and hence most adequate model fit on the final size data in the face of varying misclassification probabilities in the permissible region $[0, 0.5)$ especially when the misclassification probabilities are non zero and far apart from each other.

In chapter 8, we studied the effects of misspecification on the estimates of the stochastic SIR household epidemic model in which the epidemic model is estimated with a different infectious period distribution from the true infective period distribution of the epidemic data.

9.3 Discussion.

This work is concerned with inference for the stochastic SIR household epidemic model of [9] and [1], which are generalisations of the simple stochastic SIR epidemic model. Here, we are concerned with inference of the final size epidemic data, which may sometimes be subject to misclassification error discussed in chapters 5. These misclassification errors if ignored in our inference will lead to incorrect results of our analysis and incorrect model fit to the final size epidemic data.

It therefore means that an alternative approach to inference adopted in [1] is required in order to accommodate this scenario, by incorporating the misclassification probabilities in the modelling process. Three ways in which these errors can be handled in the modelling process were highlighted at the beginning of this thesis namely, when the misclassification probability assumed equal to zero, $\varepsilon = 0$ as in [1, 9], when they are the same, simply denoted by ε and lastly, when they are different from each other as discussed in the preamble to this work. The question then is; can the estimate of the parameters be precisely obtained under this circumstance, especially when such errors are substantially large and how do we handle the estimation problem such that precise estimates are obtained and the appropriate model fit to the final size epidemic data is identified?

The studies of the parameters and functions of the stochastic SIR household epidemic model provided insights into their properties and enhanced our understanding of their behaviours. For example the threshold parameter discussed in section 2.7 plays an important role in the occurrence of a global epidemic in the population discussed in sections 2.9, 3.7 and 4.4.

If $\lambda_L = 0$, the household structure is destroyed and the model simply reduces to the general stochastic epidemic model with household of size $n = 1$, and threshold parameter

$R_0 = \lambda_G E(T_I)$. If $\lambda_L \neq 0$ then we have the expression in equation (2.7.1).

Increasing local infection rate towards infinity, leads to increase in the threshold parameter and if the household size n is sufficiently large and $\lambda_G \neq 0$ then a global epidemic will occur in agreement with [9] and discussion in sections 3.7 and 3.11 respectively.

Reducing R_* through vaccination or otherwise also reduces the proportion of the initial susceptibles ultimately infected at the end of the epidemic, z.

Another useful community based extension of the stochastic household epidemic is that of [1] discussed in section 2.6. The [1] model extension provides computational method for the estimation of the parameters. It uses maximum likelihood algorithm derived from the assumption of independence of epidemics in each household [1], which contravenes the dependency assumption of [9]. We know this is not true as in [9] but has been employed to allow for the estimation of the parameters as in [1]. Also as observed in [9], that the event of a global epidemic in a household that did not have initial infective is distributed as that of the extended model of [1] with, $\pi = \exp(-\lambda_G E(T_I)z)$ [9].

In line with the assumption in [1] and discussion in section 2.12, we obtained the approximate likelihood function of the parameters of the final size probabilities in equation (2.12.2) and its loglikelihood function in equation (2.12.3) respectively.

The pair of the parameters (λ_L, π) and other corresponding parameters are estimated using [24] methods of generating starting values for the two dimensional model as discussed in section 4.2. The point estimates of the parameters from the Matlab programs and those of [1] discussed in section 4.3 are the same to numerical accuracy used and since the approximate likelihood function is not the true likelihood, the standard error of the estimates and their confidence intervals are inaccurate. To overcome this problem, we simulated household epidemic using the same household structure and point estimate of the parameters and then computed the mean, standard deviation and mean square error of the estimates and hence the confidence intervals. Here the standard deviation of the simulated estimates is a measure of how close the estimates are to their true parameter values.

Matlab programs to implement the estimation of the parameters are embedded in the modified version of the simhouses simulation package with subroutines which uses [24] method to

generate starting values for the estimates as discussed in section 4.2. The mean, standard deviation and mean square error are also obtained from the program function, modified simhouses simulation package. The population size should be large with adequate choice of the minimum epidemic size for the simulations as discussed and explored in sections 4.4, 4.8 and 4.9. The estimate of the model parameters are found to be unbiased with acceptable mean, small standard deviation (standard error of the estimates) and with minimum mean square error.

Appropriate choice of minimum epidemic size in simulations studies are required to allow large infections leading to unbiased estimates of the parameters as discussed in section 4.8, with illustrations in figure 4.4. The behaviour of the estimates given different minimum epidemic sizes and population sizes are explored in table 4.6 in order to provide further insights into the properties of the estimates in the face of inappropriate choice of the minimum epidemic size.

We see that it is often important to first experiment with the choice of 1, minimum epidemic size and then study the bimodal behaviour of the histogram of the number infected before choosing an appropriate cut-off between non-global and global epidemics from the simulations. The minimum epidemic size is discussed in section 4.8 and shown in figure 4.4.

From our previous studies of the two and three dimensional model estimates in sections 5.8.1, 5.9, 5.10, 5.11, 6.9 and 6.10, we found that the three dimensional model has precise and unbiased estimates and therefore the best model fit to the three dimensional final size epidemic data than the two dimensional model if the misclassification probability is not close to 0.

Programs to estimate the parameters, compute the mean, standard deviation, mean square error and plot the root mean square error of the estimates are discussed in section 5.5.

From our studies in sections 5.6 and 5.7, we found that the precision of the estimates differs for the parameter estimates, for $\varepsilon \in [0, 0.5)$. For some parameter estimates the two dimensional model outperforms the three and four dimensional models with better precision in the estimates, while for some either the three or four dimensional model is the best on the final size epidemic data. These behaviours are explored and discussed in section 5.7.

Also, the plots of the root mean square error of the estimates of the four dimensional model, are consistently stable over the misclassification probabilities region $[0, 0.2]$, compared to those of the two and the three dimensional models. These behaviours agree with those of the density histograms of the Pearson chi-square test in section 6.11 and the density histograms of the chi-square difference statistic, their empirical distribution function in section 7.10.

Here, we found that the two and three dimensional models are unable to sufficiently fit the four dimensional final size epidemic data when the misclassification probabilities are large and far apart from each other.

If the misclassification probabilities are close to each other and not 0, then the estimates of the three dimensional model are precise and unbiased including those of the four dimensional model, while those of the two dimensional model are biased, imprecise and struggled fitting the final size epidemic data. If the misclassification probabilities are close to zero, then the estimates of the two dimensional model are unbiased and precise in line with our discussion in section 5.6 including those of the three and four dimensional models. The two dimensional model is most preferred model fit to the four dimensional final size epidemic data in line with the principle of parsimony.

Considering these properties, we found the estimates from the four dimensional model to be more precise than those of two and three dimensional models on the final size epidemic data and therefore most preferred model fit when misclassification errors are known to have occur in the final size epidemic data.

If the models are misspecified on the two dimensional epidemic data, then the estimates of the three models are found to be biased with more variability around the true values from the estimates of the three and four dimensional models compared to those of the two dimensional model.

Thus, without misclassification, the more complex models are not significantly better than the two dimensional model in the presence of misspecification.

9.4 Possible Extension.

The stochastic SIR household epidemic model has been extended by different authors in various directions such as the work of [12] which extended the single type individual in [9] to multiple types (several types of individuals) in a constant population with heterogeneity in infectivity and susceptibility such that the infection rate between two individuals are dependent on the type of the transmitting and receiving individuals [12]. An Infective in class $i = 1, 2, \ldots, \mathbb{J}$ is assumed to make independent and random contacts with susceptibles in class $j = 1, 2, \ldots, \mathbb{J}$ in the population at the points of a homogeneous Poisson process having rate, $\lambda^{G}_{i,j}/N_j$, while they make contacts with susceptibles in class j within their household at the points of a homogeneous Poisson Process having rate, $\lambda^{L}_{i,j}$, similar to the community based stochastic multitype SIR household epidemic model of [1] having no global infection rate discussed in section 2.6.

The susceptibles infected from class j are infectious for period $T_I^{(j)}$ [12], after which they recover and become immune. The epidemic ceases as soon as there are no infectious individuals in the population [12].

The multitype SIR household epidemic can further be extended by introducing misclassification error in the final size epidemic data similar to this work with modification to the final size probabilities obtained from the triangular equation (2.6.1) of [1]. The likelihood function similar to that of the single type individual is the obtained by assuming independence of the epidemics in each households [1]. Hence using the maximum likelihood algorithm in [1], the local infection rates $\lambda^{L}_{i,j}$, and the vector of the escape probabilities defined as the probabilities that susceptibles of type $i = 1, \ldots, m$, avoids infection from the population, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$, using [1] notation are estimated including other epidemiological parameters similar to the single type case.

The parameter estimates of the multitype stochastic SIR household epidemic model can further be explored with different misclassification probabilities in order to provide more insights into the properties of the estimates in the face of increasing percentage noise in the final epidemic data.

If the infectious period distribution is unknown, then it is necessary to estimate the shape parameter of the Gamma infectious period distribution. For example if Gamma$(a, k/a)$ is the assumed infectious period distribution, where k is known then the shape parameter a can then estimated from the final size epidemic data.

9.5 Overall Conclusion.

We have seen from the analyses of the [1] final size epidemic data and discussion in sections 4.4, 4.8 and 4.9 respectively that the population size is required to be sufficiently large. Also required is an adequate choice of the minimum epidemic size in the simulations using the approach in figure 4.4, for the epidemic to take off.

Getting these accomplished firstly involves, simulating household epidemic with 1 as the minimum epidemic size and choosing the appropriate cut-off between the epidemics from the bimodal pattern of the histogram of the distribution of the number infected as discussed in section 4.8 and shown in figure 4.4. Simulations can then be carried out with the new choice of the minimum epidemic size after satisfactory outcome with experimenting with the minimum epidemic size of 1.

Also the level of precision of the estimates, their bias and otherwise of the three models are interpreted from the properties of their mean, standard deviation and mean square error. For example the estimates of the three models are precise and unbiased when the true final epidemic data is the two dimensional final size epidemic data, while those of the two dimensional model are biased and imprecise if the true final size epidemic data is the three and four dimensional final size data. Also if the misclassification probabilities are large and different from each other then only the estimates of the four dimensional model are precise and unbiased and therefore the preferred model fit to the final size epidemic data.

The properties of the three models are further explored using the Pearson chi-square goodness of fit test, the likelihood ratio chi-squared goodness of fit tests, the Kolmogorov-Smirnov goodness of fit test and the chi-square difference test in chapters 4, 6 and 7, in which we found that the three models sufficiently fit the two dimensional final size epidemic data, the

choice of which to use rests on the principle of parsimony which requires the use of the simplest of the three models to the final size epidemic data. Thus, the two dimensional model which requires only two parameters to be estimated is the preferred choice compared to the three and four dimensional models with three and four parameters to be estimated respectively. In general, it is often preferred fitting the two dimensional model to two dimensional final size epidemic data.

If the final size epidemic is misclassified such that the misclassification probabilities are the same then the estimates of the three and four dimensional models have precise estimates which satisfies the minimum mean square error criterion required of good estimates. The mean square error of the estimates for the three and four dimensional models tend to approximate each other for varying misclassification probabilities in the permissible region $[0, 0.5)$. The fitness of the two models is better understood from the plots of the empirical distribution of the chi-square difference statistic of the three models and their corresponding theoretical distribution, for which only three and four dimensional models are sufficient fits to three dimensional final size epidemic data.

Finally, when the misclassification probabilities are different and far apart from each other then from chapters 6 and 7, we found that only the four dimensional model is adequate on the final size epidemic data. In such situations the models tends to struggle fitting to the four dimensional model with biased estimates.

We have seen that the four dimensional model is a sufficient model fit to the four dimensional final size epidemic and the two and three dimensional final size epidemic data and therefore outperforms the two and three dimensional models. It is often useful in situation where the final size epidemic data is in doubt.

Also, with model misspecification on two dimensional epidemic data, the estimates of the three models are biased with less precision from the three and four dimensional models and more variability around the true values from the estimates of the four dimensional model.

From the chi-square difference test, we found that the three and four dimensional models are not significantly better than the two dimensional model when the epidemic data are not misclassified. If the epidemic data are misclassified then in the face of model misspecification,

the three and four dimensional models are better than the two dimensional model.

9.6 Limitation of the Study.

This work is limited to the stochastic SIR household epidemic model of [9, 11] discussed in section 2.3 and extended [1] maximum likelihood algorithm for inference to handle cases with misclassification error in the final size epidemic data. It is therefore not applicable to epidemics with different demographic settings from that of the SIR epidemic life circle and transmission pattern. For example, it cannot be applied to the SIS epidemic, which has common demographic settings with the SIR epidemic in which infectious individual recovered and immediately becomes susceptible other similar type epidemics. We have not considered infectious diseases that require birth and death demographic settings for the replenishment of the susceptibles population in order to keep the epidemic going, associated with the SI and SEI epidemics used in the study of endemic diseases.

These studies are therefore limited to the demographic settings and mixing assumptions in [9] to enable comparison of our results with those of [1] and other assumptions leading to its inference in [1] discussed in 2.6.

9.7 Recommendation.

There is the need to adjust our inference to accommodate misclassification error in the final size data, especially when they are known to have occurred, since ignoring them leads to biased estimates and choice of the wrong model. In situations when the source and methods of data collection are suspect, it may be necessary to check the level of the percentage noise if any in the final size data before embarking on inference. This is implemented using the Matlab program, falsefminsearch3(n, a, b, mat), where $a = 2$, $b = 2.05$ are the parameters of Gamma(a, b) infectious period distribution, n is the maximum household size, and mat is the matrix of the final size epidemic data. The program outputs are the maximum likelihood estimates of model parameters, λ_L, π, ε_{FN}, and ε_{FP}. If the model is two dimensional then the two misclassification probabilities will be zero or approximately zero. If it is the three

dimensional model, then the misclassification probabilities will not be zero but approximately the same. If it is the four dimensional model, then the misclassification probabilities will be different from each other. Thus, knowing the level of noise in the final size epidemic data will help determine the right model fit to the final size epidemic data.

The following are suggested procedures to follow in analysing the stochastic SIR household epidemic model, when the final size epidemic dataset is known.

1.) Run the program, falsefminsearch3(n, a, b, mat), to determine the appropriate dimension of the model and the final size epidemic data. Here the program estimates the parameters of the model including the misclassification probabilities.

2.) Calculate the Pearson chi-square and the likelihood ratio chi-square statistics, X_2, X_3, and X_4 corresponding to the three models, determine the p-values and compare the observed chi-square goodness of fit statistics for the three models with their critical values at the p-values and take decision whether to reject or not to reject the null hypothesis at the given p-values.

3.) Calculate the chi-square difference statistic, $D_{2,3}$, $D_{2,4}$, and $D_{3,4}$ using results in serial number 2.

4.) Choose the two, three or four dimensional model using the results of the analysis from $1-3$.

5.) Simulate household epidemic with 1000 repetitions using the parameter estimates and minimum epidemic size of 1 to see the bimodal behaviour of the distribution of the number infected and hence choose the appropriate minimum epidemic size which gives large infection in each simulation. Now repeat the simulations with the chosen minimum epidemic size and compute the mean, standard deviation (standard error) and the mean square error of the estimates. Also compute and examine the mean and variance of the chi-square statistics from the simulations.

Bibliography

[1] C. ADDY, I. M. LONGINI JR, AND M. HABER, *A Generalised Stochastic Model for the Analysis of Infectious Disease Final Size Data*. Biometrics, Vol. 47, No. 3, (1991), pp. 961-974.

[2] H. ANDERSSON AND T. BRITTON, *Lecture Notes in Statistics: Stochastic Epidemic Models and Their Statistical Analysis*. Springer, Verlag (2000).

[3] M. ANKER AND D. SCHAAF, *WHO Report on Global Surveillance of Epidemic-Prone Infectious Diseases*. Website: http://www.inf/emc, (2000).

[4] K. B. ANTHREYA AND P. E. NEY, *Branching Processes*, Dover books on Mathematics, Dover Publications, Inc. Mineola, New York, (2004).

[5] F. G. BALL, *The Threshold Behaviour of Epidemic Models*. Journal of Applied Probability, Vol. 20, No. 2, (1983), pp. 227-241.

[6] F. G. BALL, *A Unified Approach to the Distribution of the total size and Total Area under the Trajectory of Infection in Epidemic Models*. Advances in Applied Probability, Vol. 18, No. 2, (1986), pp. 289-310.

[7] F. BALL, *A Note on the Total Size Distribution of Epidemic Models*. Journal of Applied Probability, Vol. 23, No. 3, (1986), pp. 832-836.

[8] F. G. BALL AND D. CLANCY, *The outcome of an Epidemic Model with Several Different Types of Infectives in a Large Population*. Journal of Applied Probability, Vol. 32, No. 3, (1995), pp. 579-590.

[9] F. G. BALL, D. MOLLISON AND G. SCALIA-TOMBA, *Epdemics with Two Levels of Mixing.* Annals of Applied Probability, Vol. 7, No. 1, (1997), pp. 46-89.

[10] F. BALL AND P. DONNELLY, *Strong Approximations for Epidemic Models.* Stochastic Processes and their Application, Vol. 55, (1995), pp. 1-21.

[11] F. G. BALL AND O. D. LYNE, *Epidemics Among A Population of Households.* Mathematical Approaches for the Emerging and Reemerging Infectious Disease: Models, Methods and Theory, (The IMA Volumes in Mathematics and its Applications), Springer, Editor: Castillo-Chavez, Vol. 126, (2000), pp. 115-125.

[12] F. G. BALL AND O. D LYNE, *Stochastic Multitype SIR Epidemic Among A Population Partitioned into Households.* Advances in Applied Probability, Vol. 33, No. 1 (Mar. 2001), pp. 99-123

[13] F. G. BALL, P. O'NEILL AND J. PIKE, *Stochastic Epidemics in Structured Populations Featuring Dynamic Vaccination and Isolation.* Journal of Applied Probability, Vol. 44, No. 3 (Sept., 2007), pp. 571-585.

[14] F. G. BALL AND P. NEAL, *A general model for the stochastic SIR epidemic with two levels of mixing.* Journal of Math. Biosciences, Vol. 180, (2002), pp. 73-102.

[15] F. BALL AND P. NEAL, *Network Epidemic Models With Two Levels of Mixing.* Mathematical Biosciences, Vol. 212, (2008), pp. 69-87.

[16] F. BALL AND P. O'NEILL, *The Distribution of General Final State Random Variables for Stochastic Epidemic Models.* Journal of Applied Probability, Vol. 36, No. 2, (1999), pp. 473-491.

[17] N. T. J. BAILEY, *The Mathematical Theory of Infectious Diseases and its Applications.* Charles Griffin and Company Ltd, London, 1997.

[18] N.T.J BAILEY, *The Total Size of a General Stochastic Epidemic.* Biometrika, Vol. 40, No. 1/2, (1953), pp. 177-185.

[19] B. A. BARON, *The Effects of Misclassification on Estimation of Relative Risk.* Biometrics, Vol. 33, No. 2, (1977), pp. 414-418.

[20] M. S. BARTLETT, *Some Evolutionary Stochastic Processes.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 11, No. 2, (1949), pp. 211-229.

[21] M. S. BARTLETT, *Deterministic and Stochastic Models for Recurrent Epidemics.* Journal of Neyman (Ed), Proceedings of the Third Berkeley Symposium on Mathemaical Statistics and Probability, Vol. IV, University of California Press, (1956), pp. 81-109.

[22] , R. BARTOSZYNSKI, *Branching Processes and the Theory of Epidemic.* Fifth Berkeley symposium on Mathematical Statistics and Probability, University of California Press, Vol. 4, (1967), pp.259-269.

[23] , R. BARTOSZYNSKI, *On a Certain Model of An Epidemic.* Journal of Applicatione Mathematicae, Vol. 13, (1972), pp. 139-151.

[24] N. G. BECKER, *A stochastic Model for Interacting Population.* Journal of Applied Probability, Vol. 7, No. 3, (1970), pp. 544-564.

[25] N. G. BECKER, *Analysis of Infectious Disease Data: Monographs on Statistics and Applied Probability.* Chapman and Hall/CRC, (1989).

[26] L. BILLARD, *Factorial Moments and Probabilities for the General Stochastic Epidemic.* Journal of Applied Probability, Vol. 10, No. 2, (1973), pp. 277-288.

[27] F. BRAUER, P. VAN DEN DRIESSCHE AND J. WU (EDS.) *Mathematical Epidemiology*, Mathematical Biosciences Subseries, Springer-Verlag, Berlin (2008).

[28] D. Clancy and P. D. O'NEILL, *Exact Bayesian Inference and Model Selection for Stochastic Models of Epidemics Among a Community of Households.* Scandinavian Journal of Statistics, Vol. 34, No. 2, (2007), pp. 259-274.

[29] W.J. CONOVER, *Practical Nonparametric Statistics.* Wiley Series in Probability and Statistics:Applied Probability and Statistics Section, John Wiley and Sons Inc. (1999).

[30] D.J.DALEY AND J.GANI, *Epidemic Modelling: An Introduction*. Cambridge University Press, New York, (1999).

[31] H. E. DANIEL, *The Distribution of the Total Size of an Epidemic*. Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability Vol. IV, University of California Press, (1967).

[32] O. DIEKMANN, H. HEESTERBEEK AND T. BRITTON, *Mathematical Tools for Understanding Infectious Diseases Dynamics: Princeton Series in Theoretical and Computational Biology*. Princeton University Press, New Jersey, (2013).

[33] J. GANI, *On a Partial Differential Equation of the Epidemic Theory*. Biometrika, Vol. 52, (1965), pp.613-616.

[34] M. GREENWOOD, *On the statistical Measure of Infectiousness*. Jourrnal of Hygiene, Vol. 31, N0. 3, (1931), pp. 336-351.

[35] P. GUSTAFSON, *Measurement error and Misclassification in Statistics and Epidemiology, Impacts and Bayesian Adjustment*. Chapman and Hall/ CRC, 2009.

[36] M. E. HALLORAN, I. M. LONGINI AND C. J. STRUCHINER, *Design and Analysis of Vaccine Studies*. Statistics for Biology and Health, Springer, (2010).

[37] T. E. HARRIS, *The Theory of Branching Processes*, Prentice-Hall Inc. Englewood Cliffs, N. J. (1963).

[38] W. O. KERMACK AND A. G. MCKENDRICK, *A Contribution to the Mathematical Theory of Epidemic*. Proceedings of the Royal Society London, A, Vol. 115, (1927), pp. 700-721.

[39] C. LEFEVRE AND P. PICARD, *A Non-Standard Family of Polynomial and The Final Size Distribution of Reed-Frost Epidemic Processes*. Advances in Applied Probability, Vol.22, No.1, (1990), pp. 25-48.

[40] E. LESAFFRE, H. KUCHENHOFF, S. MWALILI, AND D. DECLERCK, *On the Estimation of the Misclassification table for finite count Data with an Application in Caries Research*. Journal of Statistical Modelling Society, Vol.9, No.2, (2009), pp. 99-118.

[41] M. LINDHOLM, *On the Time to Extinction for a Two-Type Version of Bartlett's Epidemic Model*. Mathematical Bioscience, Vol. 212 (2008), pp. 99-108.

[42] I. M. LONGINI, JR AND J.S. KOOPMAN, *Household and Community Transmission Parameters from Final Distribution of Infections in Households*. Biometrics, Vol. 38, No. 1, (1982), pp. 115-126.

[43] I. M. LONGINI, JR, J. S. KOOPMAN, A. S. MONTO, AND J. P. FOX, *Estimating Household and Community Transmission Parameters for Influenza*. American Journal of Epidemiology, Vol. 115, No. 5, (1982), pp. 736-750.

[44] A. G. MCKENDRICK, *Applications of Mathematics to Medical Problems*. Proceedings of Edinburgh Mathematical Society, Vol. 44, (1926), pp. 98-130.

[45] Mathworks, *fminsearch*. Mathworks retrieved from www.mathsworks.com/help/matlab/ref/fminse (2016).

[46] Mathworks, *Sample Kolmogorov-Smirnov Test*. Mathworks retrieved from www.mathsworks.com/help/stats/kstests.html, (2016).

[47] C. J. MODE AND C. K. SLEEMAN, *Stochastic Processes in Epidemiology*. World Scientific, Publishing Co. Pte. Ltd, (2000).

[48] M. J. MORRISSEY AND D. SPIEGELMAN, *Matrix Methods for Estimating Odds Ratios with Misclassified Exposure Data: Extensions and Comparisons*. Biometrics, Vol. 55, No. 2, (1999), pp. 338-344.

[49] P. Neal, *Efficient Likelihood-free Bayesian Computation for Household Epidemics* . Journal of Statistics and computing, Vol. 22, No.6, (2012), pp. 1239-1256.

[50] P. Neal, *A Household SIR Epidemic Model Incorporating Time of Day Effects*. Journal of Applied Probability, Vol. 53, (2016), pp. 489-501.

[51] R. PARKER, *The global HIV/AIDS Pandemic, Structural Inequalities and Politics of International Health*. Amecian Journal of Public Health, Vol. 92, No. 3, (2002), pp.343-346.

[52] S. RUSHTON AND A.J. MAUTNER, *The deterministic Model of a Simple Epidemic for more than one Community*. Imperial College, London, (1955).

[53] T. SELLKE, *On the Asymptotic Disrtibution of the Size of a Stochastic Epidemic*. Journal of Applied Probability, Vol. 20, No. 2, (1983), pp. 390-394.

[54] V. SISKIND, *A Solution of the General Stochastic Epidemic*. Biometrika, Vol. 52, 3 and 4, (1965), pp. 613-616.

[55] R. K. WATSON, *On An Epidemic in a Stratified Population*. Journal of Applied Probability, Vol. 9, (1972), pp. 659-666.

[56] P. WHITTLE, *The Outcome of a Stochastic Epidemic–A Note on Bailey's Paper*. Biometrika, Vol. 42, No. 1/2 (1955), pp.116-122.

[57] W.H.O, *Situation Report: Ebola Virus Disease 28 April 2016*. World Health Organisation, Retrieved from http://apps.who.int/ebola/ebola-situation-reports, (2016), pp. 1-10, last accessed on 13/05/2016.