Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 886-905
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Stochastic SIR Household Epidemic Model with Misclassification

Umar M. Abdulkarim

Department of Mathematics, Nasarawa State University, Keffi, Nigeria
Email: abdulmallam@yahoo.com

How to cite this paper: Abdulkarim, U.M. (2021) Stochastic SIR Household Epidemic Model with Misclassification. *Open Journal of Statistics*, **11**, 886-905.
https://doi.org/10.4236/ojs.2021.115052

Received: July 21, 2021
Accepted: October 25, 2021
Published: October 28, 2021

Abstract

In this work, we developed a theoretical framework leading to misclassification of the final size epidemic data for the stochastic SIR (Susceptible-Infective-Removed), household epidemic model, with false negative and false positive misclassification probabilities. Maximum likelihood based algorithm is then employed for its inference. We then analyzed and compared the estimates of the two dimensional model with those of the three and four dimensional models associated with misclassified final size data over arrange of theoretical parameters, local and global infection rates and corresponding proportion infected in the permissible region, away from its boundaries and misclassification probabilities. The adequacies of the three models to the final size data are examined. The four and three-dimensional models are found to outperform the two dimensional model on misclassified final size data.

Keywords

Final Size Epidemic, Infectious Period Distribution, Maximum Likelihood Estimates, Misclassification Probabilities

1. Introduction

Inference of the stochastic SIR household epidemic model without misclassification is well analyzed in [1]-[6]. The work of [1] and [7] provided maximum likelihood based algorithm for its inferences. But sometimes, the final size epidemic data is subject to misclassification error. This occurs in categorical data when the actual and recorded categories for subject differs [8] [9]. For example, the susceptibles may be wrongly be classified as infectives or an infectives wrongly classified as susceptibles. It then becomes necessary to adjust our inferences to such errors in order to get the precise parameter estimates and model that adequately fits the final size epidemic data. Using the theoretical framework developed in

this work with simulations, we explored the estimates of the parameters of the models and the adequacy of their fitness to the final size epidemic data for a range misclassification probabilities $\varepsilon_{FN}, \varepsilon_{FP} \in [0, 0.5)$. We do this by exploring the plots of the root mean square error of the estimates defined over the range of the misclassification probabilities in [0, 0.5), in order to provide clarity on the nature of their fitness to the final size epidemic data. This enables us to identify the model that adequately fits better to the final size epidemic data.

2. Material and Methods

The Model

We have assumed that the stochastic SIR household final size data of [4], is subject to misclassification error; which may be caused by susceptibles wrongly classified as infectives or infectives wrongly classified as susceptibles.

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,

$$P_{i,j}(n) = P(x+y=i \mid \text{True infect} = j, \text{household size} = n). \tag{1}$$

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,

$$\text{Bin}(j, 1-\varepsilon_{FN}) + \text{Bin}(n-j, \varepsilon_{FP}). \tag{2}$$

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

$$\text{Bin}(n-j, 1-\varepsilon_{FP}) + \text{Bin}(j, \varepsilon_{FN}). \tag{3}$$

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

$$q_{n,i} = \sum_{j=0}^{n} P(\text{Obs} = i, \text{True infect} = j, \text{household size} = n). \tag{4}$$

Since,

$$\begin{aligned} &P(\text{Obs} = i, \text{True} = j, \text{household size} = n) \\ &= P(\text{Obs} = i \mid \text{True} = j, \text{household size} = n) P(\text{True} = j). \end{aligned} \tag{5}$$

$$q_{n,i} = \sum_{j=0}^{n} P\left(x+y=i \mid \text{True infect} = j, \text{household size} = n\right) P\left(\text{True} = j\right), \quad (6)$$

where $P\left(\text{True} = j\right)$, $j = 0,1,\cdots,n$ are the final size probabilities. We can then write,

$$q_{n,i} = \sum_{j=0}^{n} P_{i,j}\left(n\right) P_j\left(n\right), i = 0,1,\cdots,n. \quad (7)$$

where

$$P_{i,j}\left(n\right) = P\left(x+y=i \mid \text{True} = j, \text{household size} = n\right). \quad (8)$$

We can generalize the expression, $P_{i,j}\left(n\right)$ for $i,j = 0,1,2,\cdots,n$ and any $r \in \mathbb{Z}_+ \le n$ using the results of $P_{0,j}\left(n\right), P_{1,j}\left(n\right), \cdots, P_{i,j}\left(n\right)$ as,

$$\begin{aligned}
P_{i,j}\left(n\right) &= \frac{j\left(j-1\right)\left(j-2\right)\cdots\left(j-i+1\right)}{r!} \varepsilon_{FN}^{j-i}\left(1-\varepsilon_{FN}\right)^{r}\left(1-\varepsilon_{FP}\right)^{n-j} \\
&+ \frac{j\left(j-1\right)\cdots\left(j-i+2\right)}{\left(i-1\right)!}\left(n-j\right)\varepsilon_{FP}\left(1-\varepsilon_{FP}\right)^{n-j-1}\varepsilon_{FN}^{j-i+1}\left(1-\varepsilon_{FN}\right)^{i-1} \\
&+ \frac{j\left(j-1\right)\left(j-2\right)\cdots\left(j-i+3\right)}{\left(i-2\right)!}\frac{\left(n-j\right)\left(n-j-1\right)}{2!}\varepsilon_{FP}^{2}\left(1-\varepsilon_{FP}\right)^{n-j-2}\varepsilon_{FN}^{j-i+2}\left(1-\varepsilon_{FN}\right)^{i-2} \\
&+ \frac{j\left(j-1\right)\left(j-2\right)\cdots\left(j-i+4\right)}{\left(r-3\right)!}\frac{\left(n-j-1\right)\left(n-j-2\right)}{3!}\varepsilon_{FP}^{3}\left(1-\varepsilon_{FP}\right)^{n-j-3}\left(1-\varepsilon_{FN}\right)^{i-3}\varepsilon_{FN}^{j-i+3} \\
&+ \cdots + \frac{\left(n-j\right)\left(n-j-1\right)\cdots\left(n-j-i+2\right)}{\left(r-1\right)!}\varepsilon_{FP}^{i-1}\left(1-\varepsilon_{FP}\right)^{n-j-i+1} j\left(1-\varepsilon_{FN}\right)\varepsilon_{FN}^{j-1} \\
&+ \frac{\left(n-j\right)\left(n-j-1\right)\cdots\left(n-j-i+1\right)}{r!}\varepsilon_{FP}^{i}\left(1-\varepsilon_{FP}\right)^{n-j-i}\varepsilon_{FN}^{j}
\end{aligned} \quad (9)$$

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

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

where $P_j\left(n\right)$ are the final size probabilities, defined as the probability of observing j infectives in a household of size n, [10] [11].

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

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

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

$$P_{r,j}\left(n\right) = \sum_{k=0}^{r}\binom{j}{r-k}\binom{n-j}{k}\varepsilon_{FN}^{j-r+k}\left(1-\varepsilon_{FN}\right)^{r-k}\varepsilon_{FP}^{k}\left(1-\varepsilon_{FP}\right)^{n-j-k} \quad (10)$$

Equations (10) is the sum of two Binomial distributions, $\text{Bin}\left(j,\left(1-\varepsilon_{FN}\right)\right)$

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,

$$P_{r,j}(n)=\sum_{k=0}^{r}\binom{j}{k}\binom{n-j}{r-k}\varepsilon_{FN}^{j-k}\left(1-\varepsilon_{FN}\right)^{k}\varepsilon_{FP}^{r-k}\left(1-\varepsilon_{FP}\right)^{n-j-r+k}. \tag{11}$$

Equation (11) is also the sum of two Binomial distributions in Equation (10) 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 (10) and (11) for $P_{r,j}(n)$ satisfies,

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

3. The Three-Dimensional Model

If the false positive and false negative misclassification probabilities are the same then Equations (2) and (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 $P_{i,j}(n), i,j=0,1,\cdots,n$ and simplified as,

$$P_{i,j}(n)=\sum_{k=0}^{i}\binom{j}{i-k}\binom{n-j}{k}\varepsilon^{j-i+2k}\left(1-\varepsilon\right)^{n-j+i-2k},\quad i,j=0,1,\cdots,n. \tag{12}$$

Alternatively, we can employ

$$P_{i,j}(n)=\sum_{k=0}^{i}\binom{j}{k}\binom{n-j}{i-k}\varepsilon^{j+i-2k}\left(1-\varepsilon\right)^{n-j-i+2k},\quad i,j=0,1,\cdots,n. \tag{13}$$

Equations (12) and (13) for $P_{i,j}(n)$ which are particular cases of Equations (10) and (11) when the misclassification probabilities are the same are made of two Binomial distributions. While Equation (12) 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 (13) 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,

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

4. Maximum Likelihood Estimation

The distribution of the final size epidemic data $x_{n,i}$ is multinomial, [12] 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, [1] [4] [13]. 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 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 $q_{n,i}\left(\lambda_L,\pi,\varepsilon_{FP},\varepsilon_{FN}\right)$.

The approximate likelihood function can be written as,

$$L\left(\lambda_L,\pi,\varepsilon_{FP},\varepsilon_{FN}\right)\propto\prod_{n=1}^{\max}\prod_{i=0}^{n}q_{n,i}\left(\lambda_L,\pi,\varepsilon_{FP},\varepsilon_{FN}\right)^{x_{n,i}}. \tag{14}$$

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,

$$\ell\left(\lambda_L,\pi,\varepsilon_{FP},\varepsilon_{FN}\right)=\sum_{n=1}^{\max}\sum_{i=0}^{n}\left(x_{n,i}\log_e\left(\sum_{j=0}^{n}P_{i,j}\left(n\right)P_j\left(n\right)\right)\right),\ i,j=0,1,\cdots,n. \tag{15}$$

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

The approximate likelihood function for the three dimensional model also has similar representation with differences in the number of parameters to be estimated.

5. Numerical Simulation and Inference 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.

Fitting the Three Models to Data from the Four Dimensional Model

We demonstrate the computational procedures of fitting the three models to four dimensional epidemic data from simulation studies and examined the behaviours of the estimates using some functions and subroutines developed for this work as,

Run the function FourDimThreeATwoSNsimhousesScatterPlotsMisspec to simulate four dimensional household epidemic data with $\text{Gamma}\left(a,b\right)$ infectious period distribution, theoretical parameters, λ_L,λ_G and $\varepsilon_{FN},\varepsilon_{FP}\in[0,0.5)$. It

then calculate the corresponding parameters of the three models with $\text{Gamma}(a,b)$ 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) $\text{LampaiD}(mat)$, provides starting values for the two dimensional model parameters, λ_L and π according to [12].

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

c) $\text{negloglik2}(x,n,a,b,mat)$, computes the negative loglikelihood function associated with the two dimensional model from the parameters of $\text{Gamma}(a,b)$ infectious period distribution, the final size epidemic data and the starting values according to [12].

d) $\text{Misclass2}(\varepsilon,n)$, computes the misclassification Probabilities associated with the three dimensional model from the misclassification probability parameter ε and maximum household size *n*.

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

f) $\text{Misclass3}(a,b,n,\pi,\lambda_L,\varepsilon)$, 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) $\text{falseMisclass2}(\varepsilon_{FN},\varepsilon_{FP},n)$, computes the misclassification probabilities associated with the four dimensional model.

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

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

j) $\text{RSTER2}(a,b,c,\lambda_L,\lambda_G,houses)$ calculates the threshold parameter, R_* from the parameters of $\text{Gamma}(a,b)$ 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 1-3** and analyzed in **Tables 1-3** respectively.

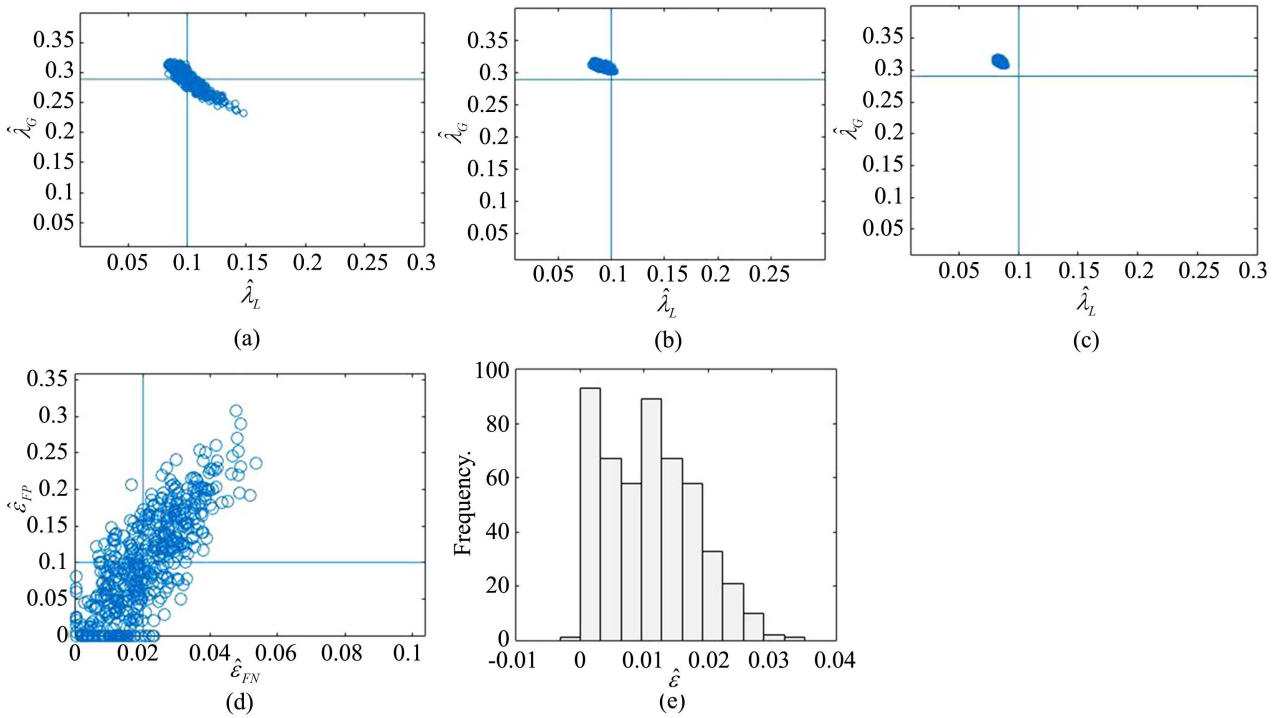

Figure 1. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.02, \varepsilon_{FP} = 0.1$. (a) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 4Dim. model; (b) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 3Dim. model; (c) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 2Dim. model; (d) Estim. $(\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP})$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

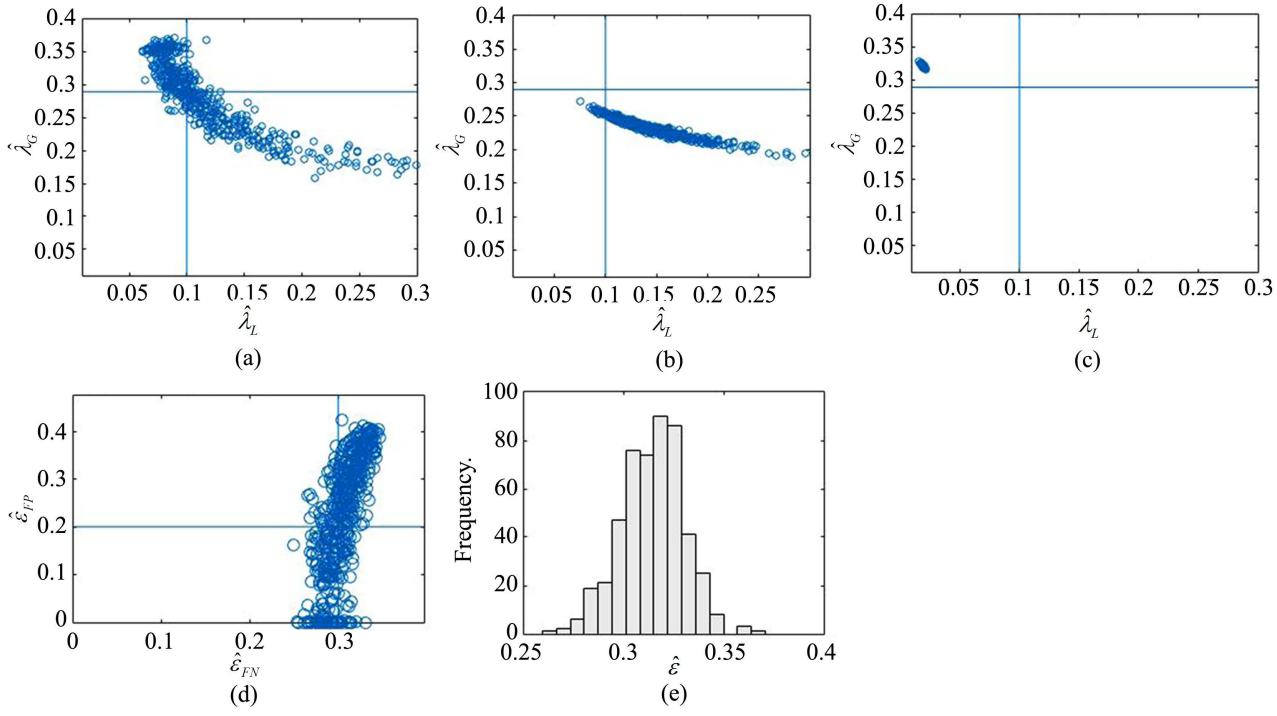

Figure 2. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.3, \varepsilon_{FP} = 0.2$. (a) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 4Dim. model; (b) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 3Dim. model; (c) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 2Dim. model; (d) Estim. $(\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP})$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

Table 1. Table of the mean of the parameter estimates of the three models.

Par. Estim.	$\varepsilon_{FN}=0.02,\varepsilon_{FP}=0.1$			$\varepsilon_{FN}=0.3,\varepsilon_{FP}=0.2$			$\varepsilon_{FN}=0.2,\varepsilon_{FP}=0.2$			Theo. Param
	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	2Dim.	3Dim.	4Dim.	
$\hat{\lambda}_L$	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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.020239	N/A	N/A	0.30444	N/A	N/A	0.20185	N/A
$\hat{\varepsilon}_{FP}$	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 2. Table of the standard deviation of the parameter estimates of the three models.

Par. Estim.	$\varepsilon_{FN}=0.02,\varepsilon_{FP}=0.1$			$\varepsilon_{FN}=0.3,\varepsilon_{FP}=0.2$			$\varepsilon_{FN}=0.2,\varepsilon_{FP}=0.2$		
	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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.011379	N/A	N/A	0.019208	N/A	N/A	0.019286
$\hat{\varepsilon}_{FP}$	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 3. Table of the root mean square error of the parameter estimates of the three models.

Par. Estim.	$\varepsilon_{FN}=0.02,\varepsilon_{FP}=0.1$.			$\varepsilon_{FN}=0.3,\varepsilon_{FP}=0.2$			$\varepsilon_{FN}=0.2,\varepsilon_{FP}=0.2$		
	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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.01137	N/A	N/A	0.019695	N/A	N/A	0.019355
$\hat{\varepsilon}_{FP}$	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

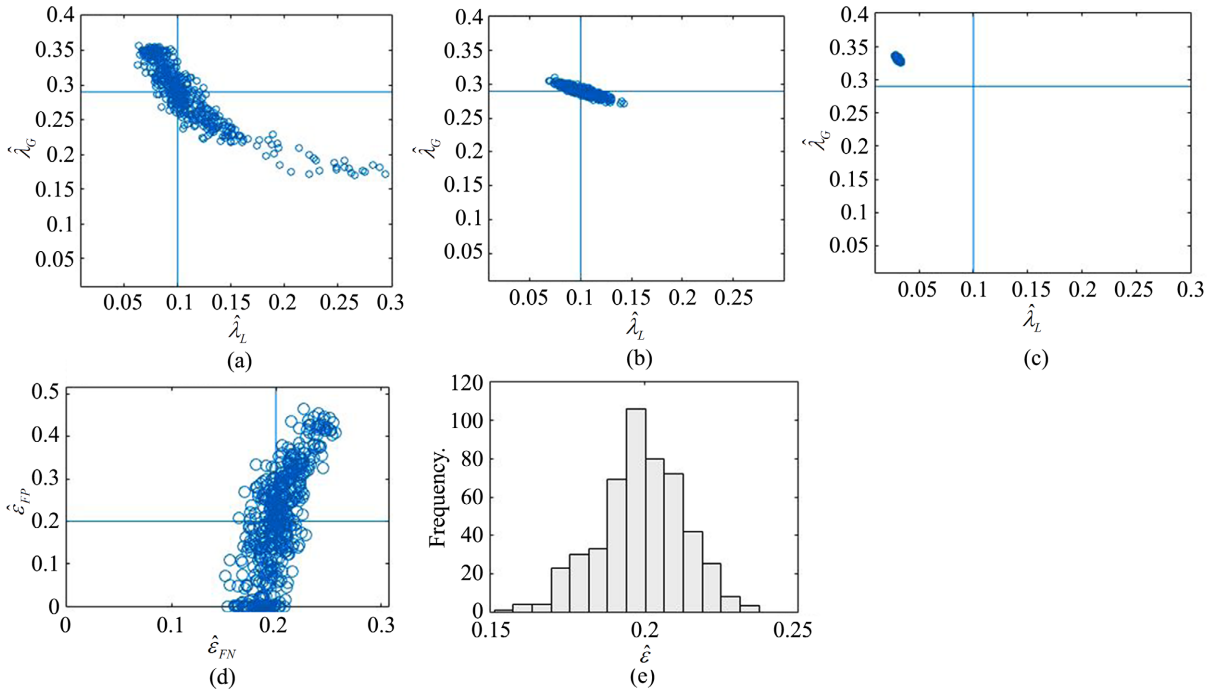

Figure 3. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon_{FN} = 0.2, \varepsilon_{FP} = 0.2$. (a) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 4Dim. model; (b) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 3Dim. model; (c) Estim. $(\hat{\lambda}_L, \hat{\lambda}_G)$: 2Dim. model; (d) Estim. $(\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP})$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

Figure 1 shows fitting the Two, Three and Four Dimensional Models to the Four Dimensional the final Size Epidemic Data, when $\varepsilon_{FN} = 0.02$, $\varepsilon_{FP} = 0.1$.

Figure 2 shows fitting the Two, Three and Four Dimensional Models to the Four Dimensional the final Size Epidemic Data, when $\varepsilon_{FN} = 0.3$, $\varepsilon_{FP} = 0.2$.

Figure 3 shows fitting the Two, Three and Four Dimensional Models to the Four Dimensional the final Size Epidemic Data, when $\varepsilon_{FN} = 0.2$, $\varepsilon_{FP} = 0.2$.

6. Comparison of the Models on the Four Dimensional Data

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, $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$ 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 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 4 (b)-(g)**.

The three dimensional model has precise estimates of λ_L for misclassification probability in $0.08 \le \varepsilon_{FP} \le 0.12$, while the four dimensional model is best

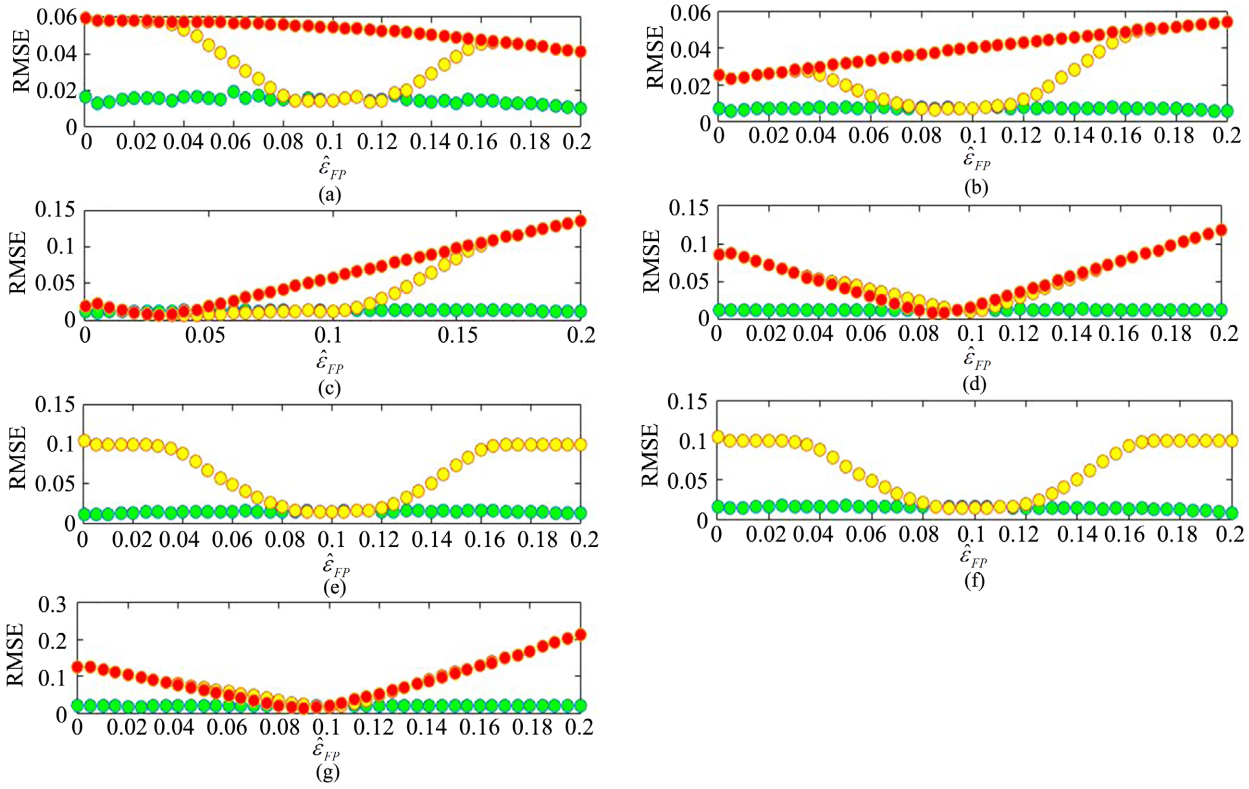

Figure 4. Plots of the root mean square error of the maximum likelihood estimates of the parameters for the three models when $\lambda_L = 0.13$, $\lambda_G = 0.17$, $\pi = 0.7423$, $z = 0.4275$, $R_* = 1.4316$. (a) Estim. of $\hat{\lambda}_L$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (b) Estim. of $\hat{\lambda}_G$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (c) Estim. of $\hat{\pi}$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (d) Estim. of z: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (e) Estim. of $\hat{\varepsilon}_{FP}$: Green = 4Dim, Yellow = 3Dim; (f) Estim. of $\hat{\varepsilon}_{FN}$: Green = 4Dim, Yellow = 3Dim; (g) Estim. of $\hat{R}_*$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim.

if $0 \le \varepsilon_{FP} \le 0.08$ and $\varepsilon_{FP} \ge 0.17$. 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 $0.08 \le \varepsilon_{FP} \le 0.01$, those of the four dimensional model are precise if, $0 \le \varepsilon_{FP} \le 0.075$ and $\varepsilon_{FP} \ge 0.115$.

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

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

In the case of the false positive misclassification probability estimates, the

three dimensional model is best, if $0.09 \le \varepsilon_{FP} \le 0.115$, while the four dimensional model is best if $0 \le \varepsilon_{FP} \le 0.085$ and $\varepsilon_{FP} \ge 0.120$ respectively.

In the case of the false negative misclassification probability, the three dimensional model is best if, $0.09 \le \varepsilon_{FP} \le 0.115$, while the four dimensional model is best if, $0 \le \varepsilon_{FP} \le 0.085$ and $\varepsilon_{FP} \ge 0.120$.

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

In summary, we see in **Figures 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 three dimensional 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.

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 $\varepsilon_{FN} = 0.2 - \varepsilon_{FP}$, $\varepsilon_{FP} \in [0, 0.2]$, 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 (a)-(g)** for a range of misclassification probabilities in $[0, 0.2]$.

From the simulation plots in **Figure 5(a)**, we see that the estimates of λ_L from the two dimensional model are driven by bias and are precise if, $\varepsilon_{FP} \le 1.975$, while the estimates of λ_L from three dimensional model are precise if, $0.050 \le \varepsilon_{FP} \le 0.165$. Those of the four dimensional model are precise if, $0 \le \varepsilon_{FP} \le 0.045$ and $\varepsilon_{FP} \ge 0.175$.

In the case of λ_G in **Figures 5(b)**, the estimates of the two dimensional model are best if, $0 \le \varepsilon_{FP} \le 0.07$, those of the three dimensional model are best if, $0.075 \le \varepsilon_{FP} \le 0.145$, while those of the four dimensional model are best if $\varepsilon_{FP} \ge 0.150$.

Also, in the case of π in **Figure 5(c)**, the estimates of the two dimensional are best if, $0.125 \le \varepsilon_{FP} \le 0.175$, those of the three dimensional model are best if, $0.07 \le \varepsilon_{FP} \le 0.120$, while those of the four dimensional model are best if, $0 \le \varepsilon_{FP} \le 0.065$ and $\varepsilon_{FP} \ge 0.18$.

In the case of z, the estimates of the two dimensional model are best if, $0.13 \le \varepsilon_{FP} \le 0.165$, those of the three dimensional model are best if, $0.065 \le \varepsilon_{FP} \le 0.125$, while those of the four dimensional model are best if,

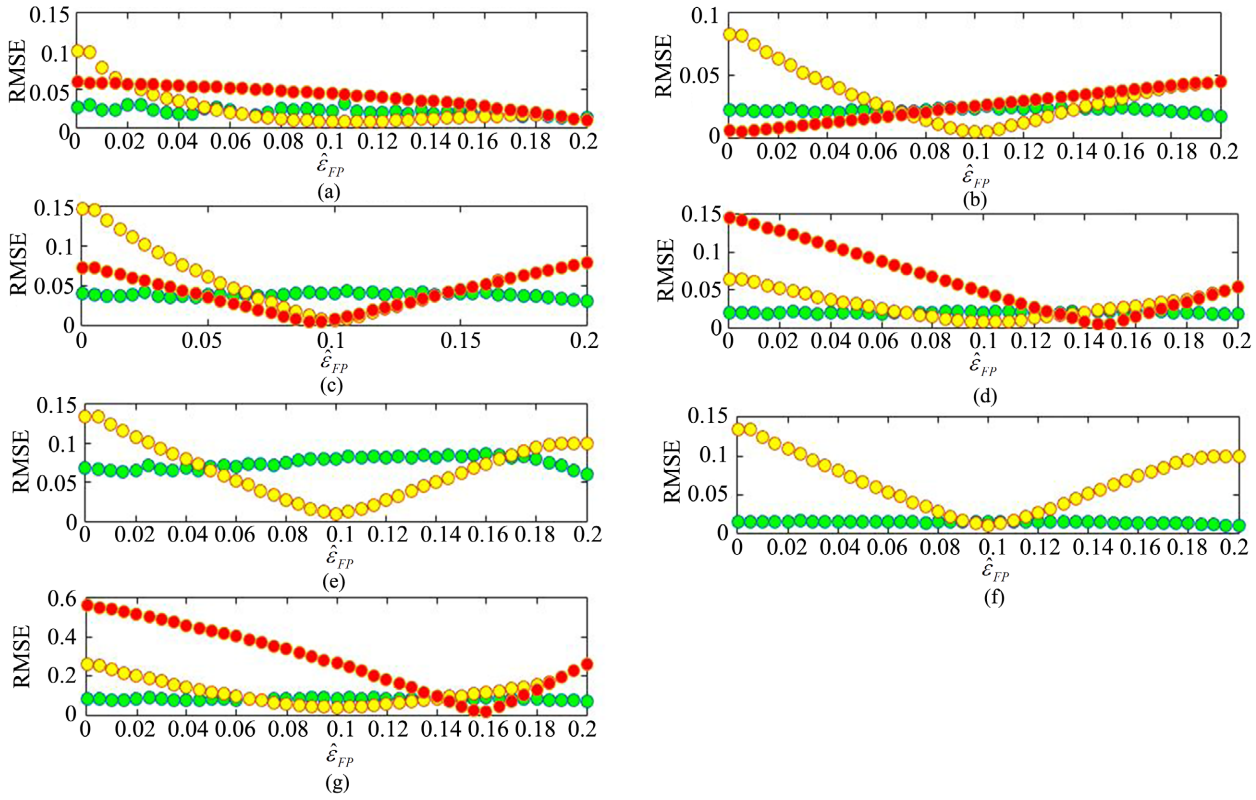

Figure 5. Plots of the root mean square error of the maximum likelihood estimates of the parameters for the three models when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $R_* = 2.2166$. (a) Estim. of $\hat{\lambda}_L$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (b) Estim. of $\hat{\lambda}_G$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (c) Estim. of $\hat{\pi}$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (d) Estim. of z: Green = 4Dim, Yellow = 3Dim; Red = 2Dim; (e) Estim. of $\hat{\varepsilon}_{FP}$: Green = 4Dim, Yellow = 3Dim; (f) Estim. of $\hat{\varepsilon}_{FN}$: Green = 4Dim, Yellow = 3Dim; (g) Estim. of $\hat{R}_*$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim.

$0 \le \varepsilon_{FP} \le 0.06$ and $\varepsilon_{FP} \ge 0.17$.

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

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

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

7. Simulation with Three Dimensional Epidemic Data

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) using the function and, subroutines developed for this work.

Figure 6 shows fitting the Two, Three and Four Dimensional Models to the Three Dimensional Simulated Final Size Epidemic Data, when $\varepsilon = 0.01$.

Figure 7 shows fitting the Two, Three and Four Dimensional Models to the Three Dimensional Simulated Final Size Epidemic Data, when $\varepsilon = 0.02$.

Figure 8 shows fitting the Two, Three and Four Dimensional Models to the Three Dimensional Simulated Final Size Epidemic Data, when $\varepsilon = 0.2$.

Table of Mean, Standard Deviation and Root Mean Square Error of the Estimates for the Two, Three and Four Dimensional Models, When $\varepsilon = 0.02, 0.02$ and $\varepsilon = 0.2$ are shown in Table 4 and Table 5.

8. Simulations and Inferences of the Two and Three Dimensional Models for $z \in [0,1]$

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] [4] [14] and 50 times its population size, minimum epidemic size of 1000, to understand the properties of the estimates. We then simulate and estimate the models parameters, compute and plot the root mean square of the estimates. Beginning with

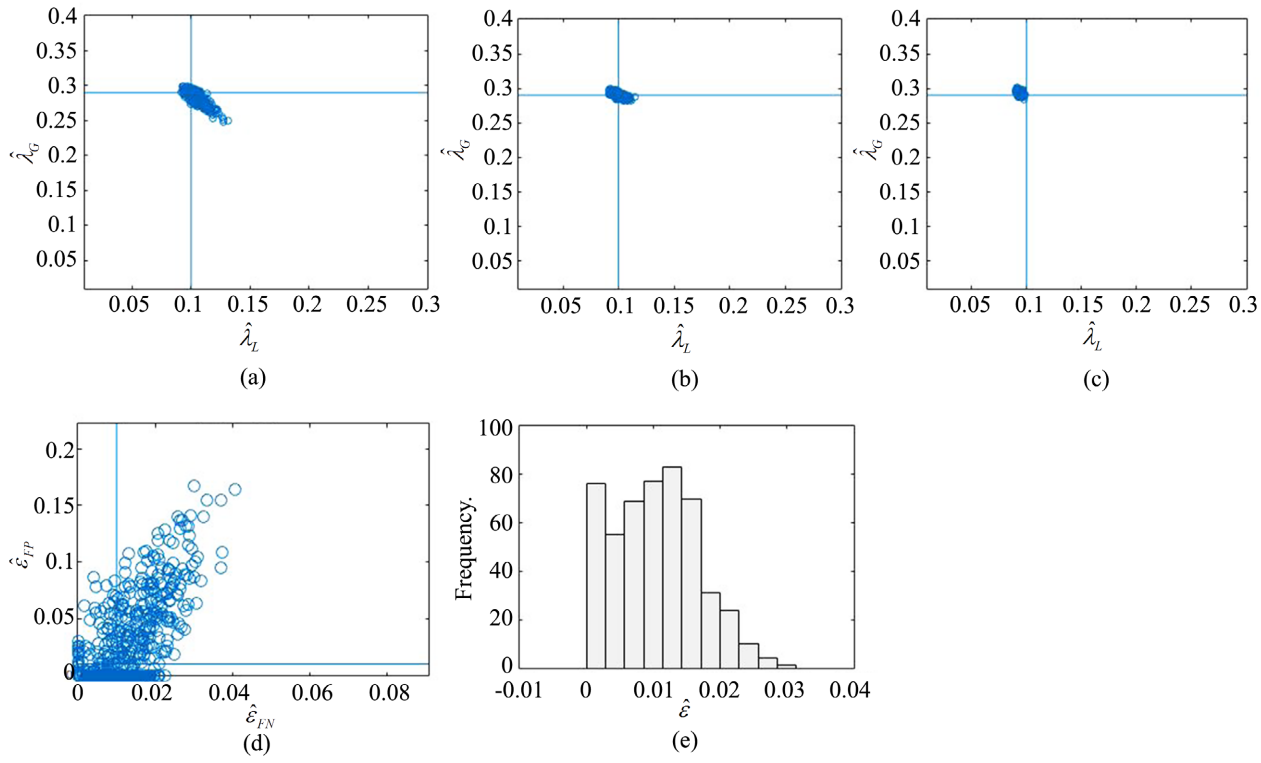

Figure 6. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.01$. (a) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 4Dim. model; (b) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 3Dim. model; (c) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 2Dim. model; (d) Estim. $\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP}$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

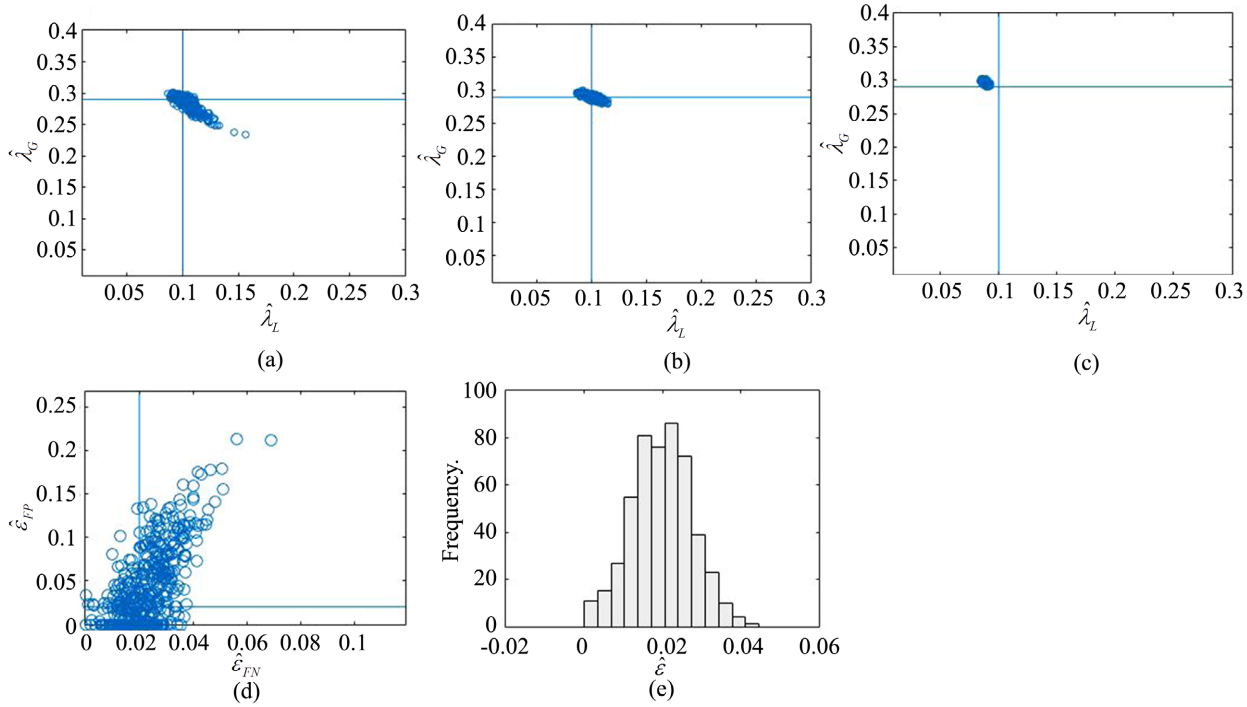

Figure 7. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.02$. (a) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 4Dim. model; (b) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 3Dim. model; (c) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 2Dim. model; (d) Estim. $\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP}$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

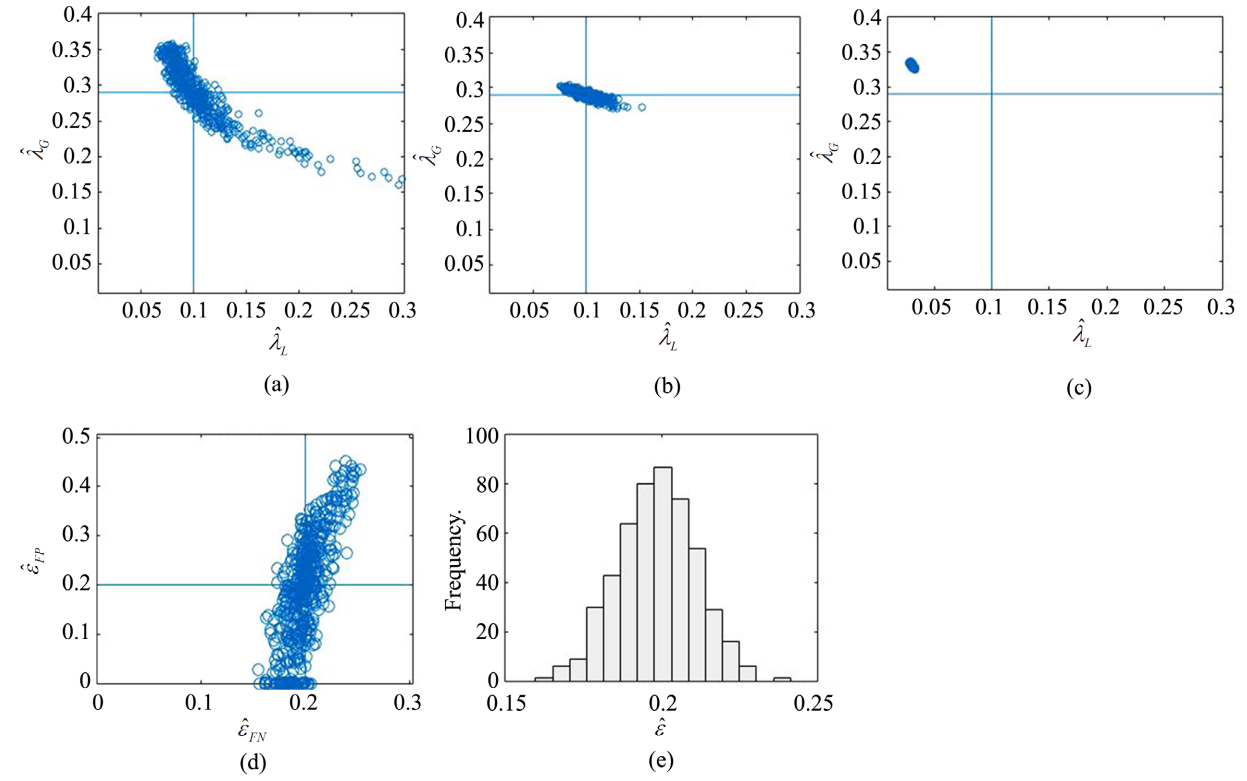

Figure 8. Plots of the estimates of (λ_L, λ_G), $(\varepsilon_{FN}, \varepsilon_{FP})$ and histogram of ε when $\varepsilon = 0.2$. (a) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 4Dim. model; (b) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 3Dim. model; (c) Estim. $\hat{\lambda}_L, \hat{\lambda}_G$: 2Dim. model; (d) Estim. $\hat{\varepsilon}_{FN}, \hat{\varepsilon}_{FP}$: 4Dim. model; (e) Hist. of $\hat{\varepsilon}$ 3Dim. model.

Table 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	Theo
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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.013024	N/A	N/A	0.022364	N/A	N/A	0.20014	N/A
$\hat{\varepsilon}_{FP}$	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. 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
$\hat{\lambda}_L$	0.0014753	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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.0079028	N/A	N/A	0.0091772	N/A	N/A	0.017635
$\hat{\varepsilon}_{FP}$	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

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 $\varepsilon \in [0, 0.1)$ (**Table 6**).

Figure 9 shows 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 10 shows Plots of the RMSE of the Parameter Estimates when $\lambda_L = 0.1$, $\lambda_G = 0.29$, $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$.

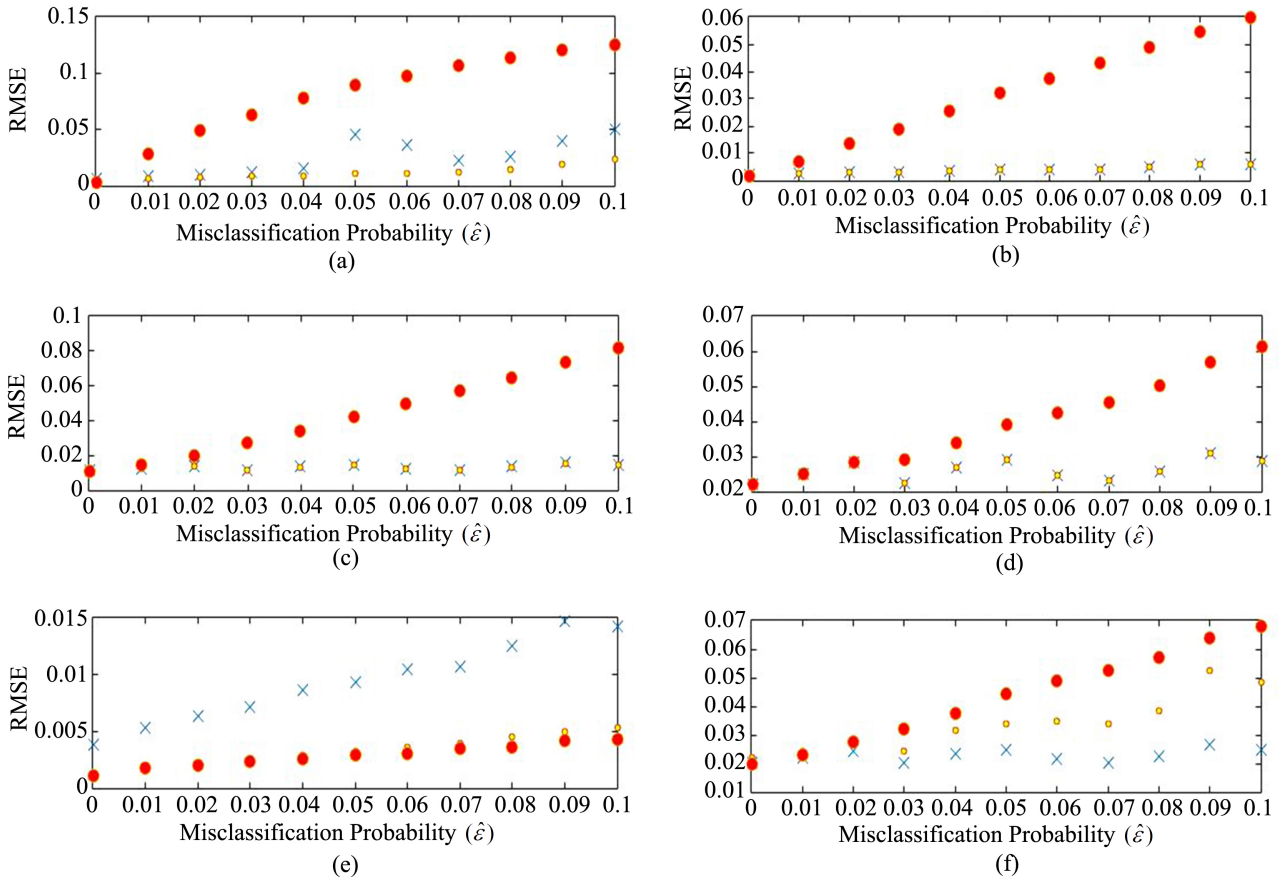

Figure 9. Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $\lambda_L = 0.2$, $\lambda_G = 0.12$, $\pi = 0.8999$, $z = 0.2144$, $R_* = 1.1653$. (a) Estim. of $\hat{\lambda}_L$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim. models; (b) Estim. of $\hat{\lambda}_G$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim models; (c) Estim. of $\hat{\pi}$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim models; (d) Estim. of $\hat{z}$: Green = 4Dim, Yellow = 2Dim; Red = 2Dim models; (e) Estim. of the Miscla. prob: Green = $\hat{\varepsilon}_{FN}$, Yellow = $\hat{\varepsilon}_{FP}$, Red = $\hat{\varepsilon}$; (f) Estim. of $\hat{R}_*$: Green = 4Dim, Yellow = 3Dim, Red = 2Dim models.

Table 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.

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
$\hat{\varepsilon}_{FN}$	N/A	N/A	0.0084544	N/A	N/A	0.0095477	N/A	N/A	0.017617
$\hat{\varepsilon}_{FP}$	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

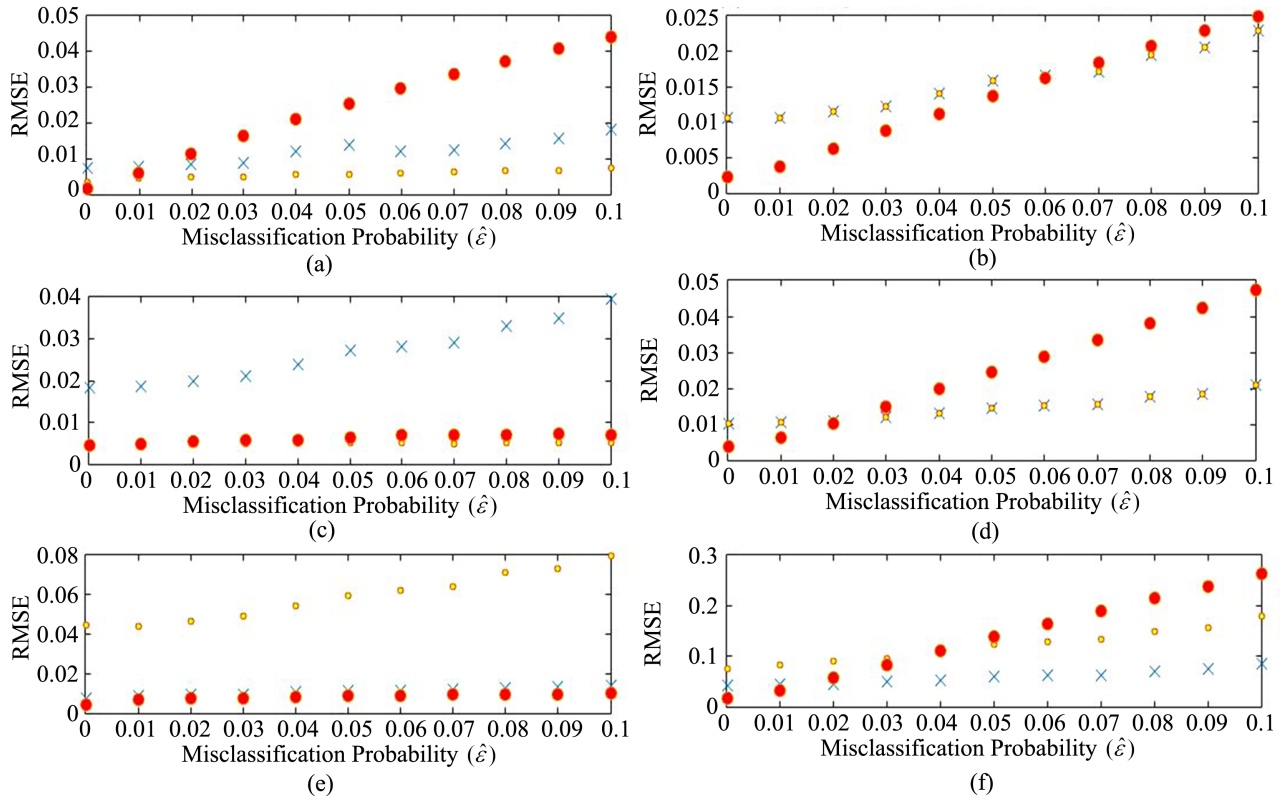

Figure 10. Plots of the RMSE estimates of λ_L for three and two dimensional optimization when $\lambda_L = 0.1$, $\lambda_G = 0.29$. $\pi = 0.4199$, $z = 0.7298$, $R_* = 2.2166$. (a) Estim. of $\hat{\lambda}_L$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim. models; (b) Estim. of $\hat{\lambda}_G$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim models; (c) Estim. of $\hat{\pi}$: Green = 4Dim, Yellow = 3Dim; Red = 2Dim models; (d) Estim. of $\hat{z}$: Green = 4Dim, Yellow = 2Dim; Red = 2Dim models; (e) Estim. of the Miscla. prob: Green = $\hat{\varepsilon}_{FN}$, Yellow = $\hat{\varepsilon}_{FP}$, Red = $\hat{\varepsilon}$; (f) Estim. of $\hat{R}_*$: Green = 4Dim, Yellow = 3Dim, Red = 2Dim models.

9. Results and Discussion

In Figures 1(a)-(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 their true values.

In Figure 2(b) and Figure 2(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.

In Figure 3(a) and Figure 3(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 Figure 3(c) are biased. The estimates of the three dimensional model are more precise than those of the two and four dimensional models.

Figures 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.

From Figure 6(c) the two dimensional models is beginning to struggle fitting to the three four dimensional data when $\varepsilon = 0.01$, while those of the three and four dimensional models are unbiased and precisely estimated as in Figure 6(a) and Figure 6(b).

Figures 5(a)-(g) provides general summary of the properties of the estimates of the three 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 examined along the vertical and horizontal axes of $[0,0.2]$ but have only chosen to present those of the former to avoid repetition.

From Figures 7(a)-(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.

From Figure 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 Figure 8(a) and Figure 8(b) are precise and unbiased as expected.

With large misclassification probability $\varepsilon = 0.2$ the three and four dimensional models are the appropriate 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.

In Figures 10(a)-(f), similar pattern of behaviour are observed except that the estimates of λ_G in Figure 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 9(c).

We see from Table 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.

Also, from the 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 $\varepsilon \in [0,0.1]$, we see that the two dimensional model is sufficient on the three dimensional final size epidemic data if ε is close to 0, while the three and four dimensional model are also sufficient model fits, if the misclassification probability is large.

The estimates of the two dimensional model are initialized according to [10], with minimum computational cost. For example from the [1] A (H3N2) Tercumseh Michigan epidemic, we found the computational time for the estimates to be 1.2 seconds, while those of the Seattle 1975-9176 B (H1N1) epidemic, [15] is 9 seconds, those of 1978-1979 A (H1N1) epidemic, [15] 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.

10. Conclusions and Suggestions

We observed that, once there is no misclassification error in the final size epidemic data, the best model fit to the two dimensional final size data is the two dimensional model. The model with smaller number of parameters is therefore 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 struggled fitting to three dimensional final size data.

With increasing ε, it becomes unreliable to use the two dimensional model. The three and four dimensional models provide good fit to the theoretical chi-square distribution in the face of increasing values of the misclassification probabilities.

Also, with large and different misclassification probabilities far apart from each other, the four dimensional model has precise estimates and therefore outperforms the two and three dimensional models on the four dimensional final size epidemic data as demonstrated.

With increasing misclassification probabilities, the two and three dimensional models struggled fitting to the four dimensional final size data, with disproportionate parameter estimates.

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 6.1.

Possible extension includes estimating the shape parameter of the Gamma infectious period distribution, if the infectious period distribution is unknown. For example if $\text{Gamma}(a, k/a)$ is the assumed infectious period distribution, where k is known, then the shape parameter a can then estimate from the final size epidemic data.

Acknowledgements

The authors will like to acknowledge Dr. Owen D. Lyne and Professor Martin Ridout for their valuable contributions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Addy, C., Longini Jr., I.M. and Haber, M. (1991) A Generalised Stochastic Model for the Analysis of Infectious Disease Final Size Data. *Biometrics*, **47**, 961-974. https://doi.org/10.2307/2532652

[2] Ball, F.G. (1983) The Threshold Behaviour of Epidemic Models. *Journal of Applied Probability*, **20**, 227-241. https://doi.org/10.2307/3213797

[3] Ball, F.G. (1986) 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*, **18**, 289-310. https://doi.org/10.2307/1427301

[4] Ball, F.G., Mollison, D. and Scalia-Tomba, G. (1997) Epidemics with Two Levels of Mixing. *Annals of Applied Probability*, **7**, 46-89. https://doi.org/10.1214/aoap/1034625252

[5] Ball, F. and Donnelly, P. (1995) Strong Approximations for Epidemic Models. *Stochastic Processes and Their Application*, **55**, 1-21. https://doi.org/10.1016/0304-4149(94)00034-Q

[6] Ball, F.G., O'Neill, P. and Pike, J. (2007) Stochastic Epidemics in Structured Populations Featuring Dynamic Vaccination and Isolation. *Journal of Applied Probability*, **44**, 571-585. https://doi.org/10.1239/jap/1189717530

[7] Clancy, D. and O'NeiIl, P.D. (2007) Exact Bayesian Inference and Model Selection for Stochastic Models of Epidemics among a Community of Households. *Scandinavian Journal of Statistics*, **34**, 259-274. https://doi.org/10.1111/j.1467-9469.2006.00522.x

[8] Baron, B.A. (1977) The Effects of Misclassification on Estimation of Relative Risk. *Biometrics*, **33**, 414-418. https://doi.org/10.2307/2529795

[9] Gustafson, P. (2009) Measurement Error and Misclassification in Statistics and Epidemiology, Impacts and Bayesian Adjustment. Chapman and Hall/CRC, London.

[10] Ball, F.G. and Neal, P. (2002) A General Model for the Stochastic SIR Epidemic with Two Levels of Mixing. *Mathematical Biosciences*, **180**, 73-102. https://doi.org/10.1016/S0025-5564(02)00125-6

[11] Neal, P. (2012) Efficient Likelihood-Free Bayesian Computation for Household Epidemics. *Journal of Statistics and Computing*, **22**, 1239-1256. https://doi.org/10.1007/s11222-010-9216-x

[12] Becker, N.G. (1970) A Stochastic Model for Interacting Population. *Journal of Applied Probability*, **7**, 544-564. https://doi.org/10.2307/3211937

[13] Ball, F.G. and Lyne, O.D. (2000) Epidemics among a Population of Households. In: Castillo-Chavez, C., Blower, S., Driessche, P., Kirschner, D. and Yakubu, A.-A., Eds., *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, Springer, Berlin, Vol. 126, 115-125. https://doi.org/10.1007/978-1-4613-0065-6_7

[14] Longini Jr., I.M. and Koopman, J.S. (1982) Household and Community Transmission Parameters from Final Distribution of Infections in Households. *Biometrics*, **38**, 115-126. https://doi.org/10.2307/2530294

[15] Neal, P. (2016) A Household SIR Epidemic Model Incorporating Time of Day Effects. *Journal of Applied Probability*, **53**, 489-501. https://doi.org/10.1017/jpr.2016.15