Statistics and Its Interface Volume 12 (2019) 415–422

Quantitative evaluation of impacts of likelihood functions on Bayesian parametric estimation of epidemic models*

Zhu Meixia, Pei Yongzhen†, Ye Ming, and Li Changguo

In epidemic modeling, the selection of likelihood function plays a crucial role on estimating model parameters and making efficient prevention strategies. Compared with the Poisson likelihood function (L_P) and normal likelihood function (L_N) based on the assumption of population homogeneity, the likelihood function, L_L, derived from Liapunov's central limit theory deals with the population heterogeneity issue that each person has a different probability of being infected. This study focuses on quantifying the performance of the three likelihood functions with particular attention paid to explore the influence of population heterogeneity on the results of parameter estimation for three epidemic models. Our results show that L_L outperforms L_P and L_N based on six sets of data, three models, and three evaluation criteria. Furthermore, L_L improves predictive capability of the three models in comparing with the prediction results of Liu et al. (2015). However, asserting the superiority of L_L for all circumstances should be cautious because the performance of the three likelihood functions are affected jointly by evaluation criteria, data sets, and the models under evaluation.

Keywords and phrases: Parameter estimation, Likelihood functions, Quantitative evaluation, Ebola.

1. INTRODUCTION

Mathematical modeling plays an important role in the field of epidemiology research. The accuracy of a model is not only a necessary foundation for determining whether the actual epidemic process can be predicted correctly, but also a major criterion in the decision-making stage of the epidemic control strategies. However, even for a well-defined mathematical model, if its parameters are unknown or ill-defined, the modeling results may be misleading. For obtaining reliable evaluation of a model, it is necessary to estimate the model's parameters first. Parameter estimation is a key issue in the process of epidemics analysis, prediction, surveillance, and control.

*The work was supported by the Natural Science Foundation of Tianjin (2017KJ092) and National Natural Science Foundation of China (11471243, 11501411).
†Corresponding author.

In the domain of epidemiological modeling, much attention has been paid to parameter estimation in the last several decades. A common way of parameter estimation is to use optimization algorithms. Based on a set of constraints and a proper initial guess, optimization is conducted by minimizing an object function that is always defined as a measure of the difference between observations and corresponding model outputs. Widely used optimization approaches have been developed, such as nonlinear least-squares fitting [1], maximum likelihood methods [2], evolutionary computation [3], genetic algorithm [4, 5] and simulated annealing [6]. However, for these methods, when the objective function is non-convex, the optimization results may heavily depend on the initial guess. In addition, many optimization approaches yield only the best parameter estimates (i.e., a point estimate), but neglect uncertainty of the estimates. Bayesian methods [7, 8] address this issue by inferring the probability distributions of parameters rather than just a point estimate.

According to the Bayes theorem, the analytical form of the posterior distribution cannot be obtained unless the denominator of the Bayes equation can be computed. To resolve this problem, sampling approaches based on Monte Carlo simulations are used to obtain a numerical approximation of the posterior distribution. Markov chain Monte Carlo (MCMC) method is one of the most popular sampling methods. To improve computational efficiency of MCMC simulation, in the last several decades, researchers have made great efforts on developing advanced sampling algorithms such as the delayed rejection adaptive Metropolis (DRAM) algorithm [9], and differential evolution adaptive Metropolis (DREAM) algorithm [10].

While MCMC has been used for estimating parameters of epidemic models, such as [11], [12] and [13], the issue of selecting an appropriate likelihood function for parameter estimation has been largely ignored. Bartlett [14] first proposed that the Poisson distribution could be used in the estimating process for most epidemic models. Daniels [15] also proved that, when population is large, the distribution of the number of infections can be approximated by Poisson distribution. Ball [16] then extended the conclusion of Bartlett and Daniels to the epidemics models that are described by random directed graph. The authors of [17] further pointed

out that the result of Daniels can be utilized when initial susceptible population size is large and infection rate per infective is constant. Since then, Poisson distribution has been widely used in epidemic studies [18, 19, 20, 21, 22]. However, McNeil [23] and Kryscio [24] suggested that a normal approximation to the number of infections can be used when the population size is large. Von Bahr and Martin Lof [25] also stated that the limit distribution of the infected population can be approximately by a normal distribution. Following these works, the normal distribution is also widely used for parameter estimation in epidemic studies, such as African cassava mosaic virus [26], plasmodium falciparum Malaria [27, 28] and 2009 influenza A(H1N1) [29].

The Poisson and normal likelihood functions, denoted as L_P and L_N, respectively, are formulated on an identical premise that each susceptible individual has the same probability of being infected at the same time. In this case, the population is regarded as being homogeneous. Otherwise, we call the population to be heterogeneous. In this paper, another likelihood function, L_L, is deduced to handle heterogeneous cases by coordinating with MCMC method. To the best of our knowledge, L_L is rarely used in epidemic models although it is has been used in other reseach fields such as hydrology field [30]. This study focused on quantifying the performance of the three likelihood functions, i.e., L_P, L_N and L_L, with a particular attention paid to explore the influence of population heterogeneity on the results of parameter estimation for epidemic models.

The remainder of this paper is organized as follows. Section 2 presents the motivation of the study, and the three likelihood functions; the comparison criteria are also introduced. The results of the comparison results of the three likelihood functions are shown in section 3. The conclusions of this study are given in Section 4.

2. BACKGROUND KNOWLEDGE

Our study bases on the following assumptions: an epidemic can be described by a mathematical model f; the true values of the parameters of f are unknown, and observed data of the epidemics are available for Bayesian estimation of the model parameters.

We denote the parameters set as a K-element vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)'$, the T-element vector of observations as $o = (o_1, o_2, ..., o_T)'$, and T-element vector of time set as $t = (t_1, t_2, ..., t_T)'$. Let $y = (y_1, y_2, ..., y_T)'$ be the unknown true infective number set, and at t_i, the relation between o_i and y_i is

$$o_i = y_i + \varepsilon_i,$$

where ε_i is the observation noise at t_i, and

$$y_i = f(t_i, \alpha),$$

where f is a model with respect to t_i and α. Capaldi [31] defines ε_i as

$$\varepsilon_i = f(t_i, \alpha)^\beta \kappa_i,$$

where $\beta \geq 0$, and $\kappa_i (i = 1, 2, ..., T)$ are independent identically distributed (i.i.d.) random variables. In this paper, we set β as 0.

The Bayesian rule [7, 8] takes the elements of the unknown true infective number set y as random variables and characterizes their uncertainties by posterior distribution

$$p(y|o) = \frac{p(y)p(o|y)}{p(o)} = \frac{p(y)L(y|o)}{\int p(y)p(o|y)dy}, \tag{1}$$

where $p(o)$ is a constant that is independent of y and is difficult to be computed out in addition to a few simple cases. An equivalent form of Eq. (1)

$$p(y|o) \propto p(y)L(y|o), \tag{2}$$

is frequently used.

Eq. (2) establishes the relation between observations o and parameters set α, because $y = f(t, \alpha)$ only depends on α when t is known. Therefore, Eq. (2) can be rewritten as

$$p(\alpha|o) \propto p(\alpha)L(\alpha|o).$$

When the prior distribution, $p(\alpha)$, is noninformative,

$$p(\alpha|o) \propto L(\alpha|o). \tag{3}$$

The Bayesian inference obeys the so-called *likelihood principle*, viz., any two posteriors that have the same likelihood functions would yield the same inference for a given sample of data. However, in practice, one rarely has the confidence of claiming that the chosen likelihood function is correct. It is thus necessary to assess the reliability of the chosen likelihood function.

2.1 Likelihood functions

Based on an assumption that the population is independent and homogenous, the status of individual j ($j = 1, 2, ...N$) at t_i can be described as a boolean random variable O_{ji}, and $O_i = \sum_{j=1}^{N} O_{ji}$ is the total observed infected number at t_i. The probability that it equals to o_i is

$$p(O_i = o_i|p_i) = C_N^{o_i}(p_i)^{o_i}(1 - p_i)^{N-o_i}, \tag{4}$$

where p_i is the probability of infection at t_i. In [15] and [23], the Poisson distribution and normal distribution are respectively taken as the approximation of Eq. (4), thus the log likelihood function $\log L(y|o)$ can be approximated by

$$L_P \quad \sum_{i=1}^{T} o_i \log f(t_i, \alpha) - f(t_i, \alpha), \text{and}$$

$$L_N \quad \sum_{i=1}^{T} -\left(\log f(t_i, \alpha) + f(t_i, \alpha) + \frac{o_i^2}{f(t_i, \alpha)}\right).$$

However, as illustrated in [20], the assumption that a large population is homogeneously mixing is unrealistic for epidemics and may lead to incorrect conclusion. By virtue of

the Lyapunov central limit theorem, we consider the impacts of population heterogeneity on parameter estimation results.

Theorem 2.1. ***(Lyapunov CLT)*** *Given a set of independent random variables X_1, X_2, ..., X_n, where $E(X_i) = \mu_i$ and $var(X_i) = \sigma_i < \infty$ $(i = 1, 2, ..., n)$, then*

$$\Phi(x) = \lim_{n\to\infty} p\left(\frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum \sigma_i^2}} \le x\right).$$

On the assumption that the susceptible population is independent but heterogenous, i.e., at t_i each individual j $(j = 1, 2, ...N)$ has a probability p_{ji} of infection, then the observed status of j at t_i could be defined as a boolean random variable O_{ji}, whose expectation μ_{ji} and variance σ_{ji}^2 are respectively p_{ji} and $p_{ji}q_{ji}$, where $q_{ji} = 1 - p_{ji}$. $O_i = \sum_{j=1}^{N} O_{ji}$ is the number of observed infected individuals at t_i. By the Lyapunov CLT, we get that

$$p(O_i = o_i | Y_i = y_i) = \frac{1}{\sqrt{2\pi \sum_{j=1}^{N} p_{ji}q_{ji}}} e^{-\frac{(o_i - y_i)^2}{2\sum_{j=1}^{N} p_{ji}q_{ji}}}.$$

Then, the log-likelihood function is approximated to

$$L_L^* \quad -\sum_{i=1}^{T} \left[\log\left(\sum_{j=1}^{N} p_{ji}q_{ji}\right) + \frac{(o_i - f(t_i, \alpha))^2}{\sum_{j=1}^{N} p_{ji}q_{ji}} \right].$$

Based on Box and Tiao [32], we have

$$L_L \quad \log L(y|o) \propto -\sum_{i=1}^{T} (o_i - f(t_i, \alpha))^2, \tag{5}$$

when the prior distribution of $p_{ji}q_{ji}$ is noninformative.

L_L has been used as an alternative likelihood function of the MCMC technique, DREAM(ZS) [10], in estimating parameters of hydrological models in [30, 33]. In the domain of epidemics, L_L is rarely used as a likelihood function in spite of its flexibility to handle population heterogeneity. Comparing with the frequently-utilized likelihood functions L_P and L_N, it is unknown to what extent the likelihood function L_L affects the accuracy of the uncertainty quantification in epidemic models. Therefore, we compare the performance of L_P, L_N, and L_L through three representative models.

2.2 Comparison criteria

The residual sum of squares (RSS), which is utilized in many researches to evaluate the performance of parameter estimation and model selection, is defined as

$$RSS = \sum_{i=1}^{T} (o_i - f(t_i, \hat{\alpha}))^2, \tag{6}$$

where $\hat{\alpha}$ is a vector of the parameters whose value are estimated based on the observed data set o. Smaller RSS indicates lower disparities between the observations and corresponding model simulations. Yet, as stated by Larson [34], running and evaluating the statistical performance of a method on the same data would lead to over-fitting. Evaluating the output of a model based on new data can result in more reliable conclusions of the model's performance [35], we use another statistical index, that is, residual sum of squares of predicted data (RSS_p), defined as

$$RSS_p = \sum_{j=k}^{T} (o_i - f(t_i, \tilde{\alpha}))^2, \tag{7}$$

where $\tilde{\alpha}$ denotes the parameter values estimated based on samples $o_1, ..., o_{j-1}$. An estimator is more reliable than another estimator if it has a smaller RSS_p.

In practice, Eq. (6) and Eq. (7) are often implemented by the cross-validation (CV) approach. In a m-fold CV, data set D is splitted into k mutually exclusive subsets D_1, D_2, ..., D_m of approximately equal size [36]. At step $i \in \{1, 2, ..., m\}$, the training set is D/D_i and the validating set is D_i. This leads to the mean residual sum of squares (MRSS) defined as

$$MRSS = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{|D_i|} (D_i(j) - f(t_j, \hat{\alpha}^{-i}))^2, \tag{8}$$

where $D_i(j)$ is the j_{th} data of D_i, $\hat{\alpha}^{-i}$ is the parameters' estimating result based on training data D/D_i, and t_j is the time stamp of $D_i(j)$. Kohavi [36] pointed out that 10-fold CV may be the best method even when computation power allows for more folds.

3. DATA AND RESULTS

The performance of the three likelihood functions was first analyzed by using the "true" data generated by using "true" parameter values and a "true" model. This allows us inspecting each likelihood function to investigate how well it can retrieve the "true" parameter values. The advantage of using simulating data instead of actual data is that, since the true parameter value is known, we can directly evaluate the performance of parameter estimation and model prediction. The disadvantage is that it is difficult to ensure the general applicability of the conclusions obtained from simulation data, since these conclusions may be influenced by the model's structure and the simulated data. This problem can be resolved by examining the likelihood functions' performance with enough models and data, which however is beyond the scope of this study. One less-than-ideal alternative option is to consider their performance by a set of classical models and several different datasets.

In this paper, we consider three classical epidemic models, i.e., the Logistic, Gompertz and Richards model, and three simulating data sets (D_L, D_R and D_G) that are respectively produced by the three models (Eq. (9)–Eq. (11)). The simulated data are post-processed by adding white noise. After evaluating the three models using the simulated data,

the real data of Ebola epidemics in Africa is used to further evaluate the performance of the likelihood functions to draw conclusions that are more realistic.

3.1 Simulated data and results of parameter estimation

The three models are given below:

- Logistic model [37]: $Y_L'(i) = rY_L(i)\left(1 - \frac{Y_L(i)}{K}\right)$. Its analytical solution is

$$Y_L(i) = \frac{Ky_0}{y_0 - (y_0 - K)e^{-ri}}, \tag{9}$$

  where r is per capita growth rate of the infected cases, K is the carrying capacity and y_0 is the initial number of infected cases.
- Richards model [38]: $Y_R'(i) = rY_R(i)\left(1 - (\frac{Y_R(i)}{K})^q\right)$. It is a variant of the Logistic model, and incorporates an exponential term q to depict the exponent of deviation from the standard logistic curve. Its analytical solution is

$$Y_R(i) = K\left(1 + \left(\frac{y_0}{K}\right)^{-q} e^{-rqi}\right)^{-\frac{1}{q}}. \tag{10}$$

- Gompertz model [39]: $Y_G'(i) = rY_G(i)\ln\left(\frac{K}{Y_G(i)}\right)$. Its analytical solution is

$$Y_G(i) = e^{(\ln(y_0)e^{-ri} + (1+\ln K)(1-e^{-ri}))}. \tag{11}$$

We assume that true values of r, q, K and y_0 are respectively 0.6, 4, 40000 and 50, and the observation interval is measured by day and the observation period is 120 days. The simulated data D_c are acquired by

$$D_c(i) = Y_c(i) + \varepsilon_i,$$

where $Y_c(i)$ is the true data got by Eq. (9) to Eq. (11), $\varepsilon_i \sim N(0, 20)$ and $c \in \{L, R, G\}$.

Before starting the parameter estimation, we review the relationships between the nature of populations and the three likelihood functions. We associate a model with an epidemic and represent the cumulative number of infections at t_i by the model's solution $y_i = f(t_i, \alpha)$. If we take the susceptible population as i.i.d., then $y_i = Np_i$, where N is the number of susceptible population, p_i is the individual's probability of being infected at t_i. In this case, the probability that the observed infected number at t_i equals to o_i is calculated by Eq. (4). According to [15] and [23], Eq. (4) can be approximated by the Poisson distribution and normal distribution, based on which L_P and L_N are the likelihood functions when a population is assumed to be homogeneous. When population is heterogeneous, $y_i = Np_i$ is wrong, since the probability of being infected is different for different person. It is more appropriate to denote y_i as

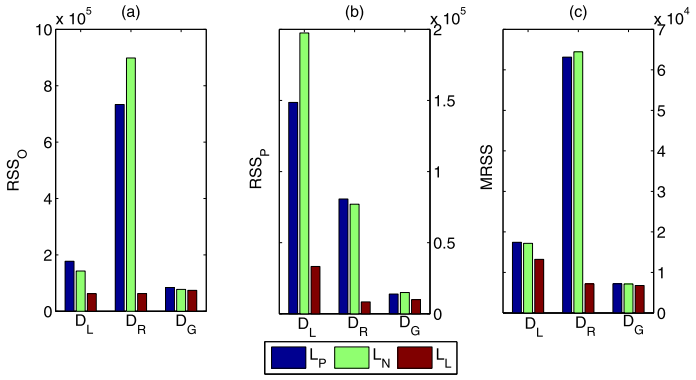

Figure 1. Results of the performance comparison among the likelihood functions L_P, L_N and L_L by RSS, RSS_p and $MRSS$.

$\sum_{j=1}^{N} p_{ji}$, and by the Lyapunov's central limit theorem, we can deduce the likelihood function L_L, namely, Eq. (5) to deal with the heterogenous issue.

The MCMC simulation for parameter estimation and uncertainty quantification is conducted by running three chains. Latin hypercube sampling method is used to generate random parameter values. Uniform distribution is taken as the prior parameter distributions. The scale reduction factor ($\hat{R}$) [40] is used to monitor the convergence of the MCMC simulation, and convergence is achieved when $\hat{R} \leq 1.2$. We find that the convergence goal is fulfilled in each data set for all three likelihood functions. To take D_G and L_N as an example, the convergence result that is shown in Figure S1 in Supplementary Material http://intlpress.com/site/pub/files/_supp/sii/2019/0012/0003/SII-2019-0012-0003-s001.pdf indicates all the MCMC chains achieve convergence after 9000 runs. The estimating results of the parameters are listed in the supplementary Material (Table S1 in Supplementary Material), and used as an alternative explanation of the convergence performance.

As for RSS and RSS_p, we first produce the predicted data sets $D_L'(t)$, $D_R'(t)$ and $D_G'(t)$ for $t \in \{121, 122, ..., 140\}$ in the same way as $D_L(t)$, $D_R(t)$, and $D_G(t)$, then by Eq. (6) and Eq. (7), the value of RSS and RSS_p are calculated and shown in Figure 1(a) and Figure 1(b). As shown in Figure 1(a), if RSS is used as the evaluation criterion, the performance of L_L is the best for data sets D_L and D_R. For data set D_G, the performance of the three likelihood functions are similar. When RSS_p is used as the evaluation criterion, L_L evidently performs best as shown in Figure 1(b). In addition, as for data sets D_R, D_L', and D_R', L_L is superior to L_P and L_N. Therefore, the performance of a likelihood function may rely on the data set when the evaluation criterion is fixed. On the other hand, the performance of different likelihood functions may be different for the same data set when the performance is judged based on different criteria. Therefore, it is not advisable to draw

conclusions on the results of parameter estimation without carefully comparing the performance of alternative likelihood functions.

We also use 10-fold cross validation to check the performance of the likelihood functions on different data sets. The MRSSs are shown in Figure 1(c). For all the data sets, the MRSSs values of L_P and L_N are nearly identical. For D_R, the difference between L_P (or L_N) and L_L is very large, while the difference is small for D_G. For D_L, the difference is moderate. Although L_L gives the least MRSS in all data sets when $MRSS$ is used as criterion, the differences in the performance of the three functions vary.

Predictive uncertainty of the likelihood functions is quantified by using the 95% confidence intervals estimated from the data sets D_L, D_R and D_G. The confidence intervals of the differences between estimations and observations are presented in Figure S2(a)–Figure S2(c) in Supplementary Material. As for D_L, the intervals induced by all three likelihood functions are too narrow to cover any the observations before time 60. With regard to D_R, the likelihood functions are also not wide enough to cover the observations before time 50. The intervals induced by L_L is narrower than the intervals of the other two functions, and cannot cover the observations until time 70. D_G is much better than D_L and D_R, since its observations are all covered by the intervals induced through all three likelihood functions. The mean predictions, i.e., the center of the intervals, are shown in Figure S2(d)–Figure S2(f) in Supplementary Material. For data sets D_L and D_R, the mean predictions obtained by employing L_P and L_N are closer to the observations than those obtained by employing L_L. However, for data set D_G, the mean predictions obtained by L_P and L_N are more biased than those obtained by using L_L. These results manifest that L_P and L_N have the similar predictions on all three data sets, while L_L gives different predictions. Therefore, it is necessary to consider all three likelihood functions for quantifying predictive uncertainty of epidemics modeling, even though the corresponding MCMC simulations are computationally expensive.

3.2 Ebola data and results

Ebola, which is characterized by diarrhea, fever, and severe vomiting, has a high fatality rate and has been classified as a serious epidemic by the World Health Organization (WHO). The last outbreak of Ebola emerged in Africa in 2014. The outbreak was first identified in Guinea, then it spreaded to the neighboring areas of Liberia and Sierra Leone. The observed data sets of Guinea, Liberia and Sirrea Leone are available at http://www.who.int/ebola/en/. In order to compare the performance of the three likelihood functions in real epidemic outbreaks, we used the accumulated infective number that are gathered from Guinea, Liberia, Sierra Leone as observations. In addition, there is no report on the Ebola in other West Africa countries, so we use the three countries' total time series data of the three countries as the observed data set of the West Africa area.

The epidemic dynamics of the four areas are simulated by the Logistic, Richards and Gompertz model. The model parameters are estimated by using the likelihood functions, L_P, L_N and L_L. As shown in Table S2 in Supplementary Material, the RSS induced by L_L is the smallest. Therefore, if RSS is used as a comparison criterion, L_L is the likelihood function that achieves better fitting in each model.

We further considered the prediction capabilities of the likelihood functions. From the web http://www.who.int/ebola/en/, we obtained the reported data from May 4, 2015, to June 12, 2015 as the prediction data. Figure S3 to Figure S9 in Supplementary Material represent the predicting results of the different likelihood functions. Furthermore, in order to survey the effectiveness of the likelihood function L_L, we also cite the result of Liu [41] as a comparison object. For Ebola, by means of the likelihood function L_P and adaptive MCMC method, Liu estimated parameters of three models, i.e., Logistic, Gompertz and Richards models. With purpose of emphasizing the importance of model selection in parameter estimating process, Liu got the following key findings:

- As for Guinea and the West Africa area, Logistic model is slightly better than Richards model, but they are both better than Gompertz model. Therefore, in predicting stage, Liu used Logistic model to generate predicted number of infections.
- As for Liberia, Richards model is the best. The predicting results of the Logistic and Gompertz models are very close, but they are not as good as that of the Richards model. Therefore, he used Richards model to generate predicting data of infections of Liberia.
- Richards model is used to generate Sierra Leone's predicted number of infected persons, because its behavior is far better than those of the other models where Gompertz model is worse and Logistic model is the worst.

For clarity, Table S3 in the Supplementary Material is used to describe the model selecting result of Liu.

As for Guinea, the predicted result of the Logistic model with likelihood L_L (Figure S3(a)) is significantly better than those of the others. For the Gompertz model (Figure S3(b)), the performance of all the three likelihood functions are not as good as that of Liu, and it is because that Liu used model selection method and chose the Logistic model as the best one for Guinea. As for the Richards model, Figure S3(c) shows that all the three likelihood functions underestimate the incidence of Ebola. But comparing with the result of Liu which overestimated the incidence, Figure S4 in Supplementary Material shows that the absolute error of L_L is smaller than that of Liu for the Richards model. As a summary, for Guinea, the L_L likelihood function is much better than L_P and L_N for all the three models, and the predicting capability of L_L is superior to that of Liu for the Logistic or

Richards model. Yet, the prediction results obtained by all three likelihood functions are inferior to those of Liu for the Gompertz model, because Liu used model selection method and found that the Gompertz model is not suitable for the data obtained from Guinea.

The predicted results of Liberia by all the three likelihood functions and epidemic models are illustrated in Figure S5 in Supplementary Material. Liu used the Richards model to obtain the predicted infective number. For the Logistic model, all the predictions underestimate the incidence of Ebola, but Liu obtained better prediction because he used the preferred Richards model. For the Gompertz model, the predictions of all the three likelihood functions overestimate the Ebola incidence, while Liu underestimated it. But as has shown in Figure S6 in Supplementary Material, the absolute difference ratio of L_L is $2\% - 3.5\%$, which is smaller than that of Liu (nearly 8%). This finding is controversial to that of Liu who has claimed the Richards model is the best for the data from Liberia. Therefore, for the Liberia data, when all the three likelihood functions are compared, the performance of L_L is the best. Moreover, the fact that L_L gives better performance for the Gompertz model than Liu manifests the impact of likelihood function on model selection.

Figure S7 in Supplementary Material illustrates the prediction result of Sierra Leone. Through model selection, Liu concluded that the Richards model is the best, and the Logistic and Gompertz models are not recommended. This conclusion is confirmed in our study for the Logistic (Figure S7(a)) and Gompertz (Figure S7(b)) models, in that the performance of all three likelihood functions is inferior to that of Liu. When we also use the Richards model, likelihood function L_L gives more accurate results than L_P and L_N, but it is not as good as that of Liu.

The predicting result of West Africa is illustrated in Figure S8 in Supplementary Material. Since Liu showed that Richards model is the best, none of the performance of the three likelihood functions is superior to that of Liu when we take the Logistic model as our proposed model. However, when the three likelihood functions are used for comparison, we find that L_L and L_N are better than L_P. For the Gompertz model, the predictions of all the three likelihood functions overestimate the incidence of Ebola, but the absolute difference ratio of L_L is nearly identical to that of Liu (as shown in Figure S9 in Supplementary Material). Therefore, the conclusion of Liu that Richards model is preferable for West Africa needs to be examined, since the predictive ability of the Gompertz model with likelihood function L_L is not inferior to it. Likelihood function L_L performs better than the other two likelihood functions and Liu when we adopt Richards model.

The MRSSs of the four areas for the three models and the three likelihood functions are listed in Table S4 in Supplementary Material. The performance of different likelihood functions is compared when the epidemics model is fixed. In each model and area, the MRSS obtained by L_L is smaller than or similar to the others.

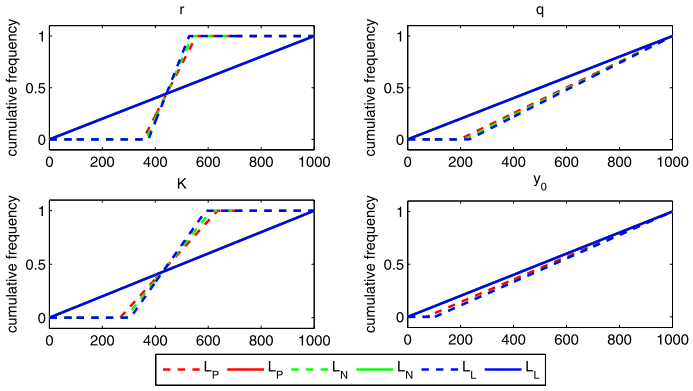

Figure 2. MPSA results of the Richards model in Sierra Leone with likelihood functions L_P, L_N and L_L. Acceptable and unacceptable cases are denoted as dashed and solid lines respectively. Sensitivity degree of each parameters are represented by the separation extent between two cases.

3.3 Parameter sensitivity analysis

Using the Richards model as an example, we identified the relatively sensitive and hence important parameters by employing the multiparametric sensitivity analysis (MPSA) method of [42].

The parameter sensitivity is evaluated by comparing two 'cumulative frequency' distributions that are associated with the acceptable and the unacceptable cases. If the two distributions are similar, the parameter is identified as insensitive, otherwise sensitive. For the Sierra Leone data, the cumulative frequency distributions of the unacceptable and acceptable cases are shown in Figure 2 for the Richards model. The MPSA results identified that the sensitivity order of the parameters is $r > K > y_0 > q$ for all likelihood functions. The consequence of per capita growth rate of the infected cases (r) is the most important parameter (Figure 3(a)), since a critical index in epidemics field, namely, the basic reproduction number R_0 is calculated by formula $R_0 = exp(rT)$ [41], where T is the the duration of the infectious period. The result that q is the most insensitive parameter is also reasonable in accordance with the model selection consequence of [41]. As for the other countries whose MPSA ordering results are displayed in Table S5 in Supplementary Material, their MPSA results are not shown in graphical format because of space limit. The sensitivity ordering is the same for all the three likelihood functions. Therefore, L_L is more reliable since it obtains less residual sum of squares in the parametric estimating process and better verification results in the prediction process.

The orders of sensitivity are also analyzed by computing the acceptance rate under the same range of parameters (Figure 3). A large acceptance rate of a parameter means its low sensitivity. Its results are same as those of the MPSA approach.

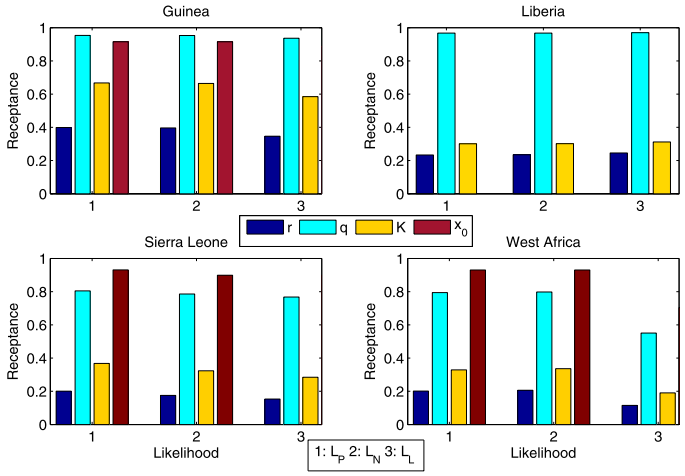

Figure 3. Acceptance rates when the parameters are set in the same range as MPSA.

4. DISCUSSION

In epidemiological area, the selection of likelihood function as well as its impact on parameter estimation plays a crucial role on making efficient preventing strategies. In this paper, three likelihood functions denoted by L_P, L_N and L_L, are respectively employed. Based on an assumption that the population is homogeneous, L_P and L_N, have served well in the epidemic parameter estimation process. What we should note here is that the homogeneous assumption is ideal and each person may has different probability of being infected, i.e., the population is heterogeneous. L_L that deals with the population heterogeneity issue is active in the hydrology field [30], whereas it is rarely used in the epidemic area. So we intend to explore the influence of population heterogeneity on the results of parameter estimation for epidemic models by quantifying the performance of L_P, L_N and L_L on three traditional models, and the primary findings are shown as follows:

- On all data sets, the estimations stemmed from likelihood functions L_P and L_N are similar, whereas they are different from those obtained by employing L_L. Furthermore, the performance of L_L is superior to those of L_P and L_N in view of all criteria.
- Likelihood functions can exert imperative impacts on model selection. By employing L_P as the likelihood function of the model selection process, Liu [41] explores the parameter values and claims that Gompertz model may not be a suitable candidate for describing the Ebola data. However, we find that the formerly 'unrealistic' Gompertz model becomes preferable when L_L is taken as the likelihood function. This illuminates that appropriate likelihood function should also be deliberated to improve model's predicting capability in the model selecting process.

In summary, the outstanding performance of L_L reinforces the indispensability of considering population heterogeneity in the parameter estimating process for epidemic models. However, it is unadvisable to assert the superiority of L_L over L_P and L_N, since the comparison results obtained from a limited number of models, data sets and criteria can't be generalized to all probable scenarios. The appropriateness of L_L should be further considered on other epidemic models such as West Nile Fever, Nipah virus and Ehrlichiosis, and selecting an appropriate likelihood function is also worthy of attention for parameter estimating and model selecting of deadly epidemics such as Ebola.

ACKNOWLEDGEMENT

The authors would like to thank the two reviewers for their valuable and useful comments that have helped us to improve this paper. We are also grateful to Co-Editor Ming-Hui Chen for his great help in reformatting this paper.

Received 3 April 2018

REFERENCES

[1] Chick, S. E., Koopman, J. S., Inferring infection transmission parameters that influence water treatment decisions, Management Science 49 (7) (2003) 920–935.

[2] Clark, M., Vynnycky, E., The use of maximum likelihood methods to estimate the risk of tuberculous infection and disease in a canadian first nations population, International Journal of Epidemiology 33 (3) (2004) 477–84.

[3] Jara, A., Martinez, R., Vigueras, D., Sanchez, G., Attribute selection by multiobjective evolutionary computation applied to mortality from infection in severe burns patients (2011) 467–471.

[4] Patel, R., Longini, I. M., Jr., Halloran, M. E., Finding optimal vaccination strategies for pandemic influenza using genetic algorithms, Journal of Theoretical Biology 234 (2) (2005) 201–212. MR2135192

[5] Liao, J. Q., Hu, X. B., Wang, M., Leeson, M. S., Epidemic modelling by ripple-spreading network and genetic algorithm, Mathematical Problems in Engineering, 2013 (2013-11-3) 2013 (9) (2013) 841–860. MR3092981

[6] Suman, B., Kumar, P., A survey of simulated annealing as a tool for single and multiobjective optimization, Journal of the Operational Research Society 57 (10) (2006) 1143–1160.

[7] Huang, X., Lambert, S., Lau, C., Soares, M. R. J., Marquess, J., Rajmokan, M., Milinovich, G., Hu, W., Assessing the social and environmental determinants of pertussis epidemics in queensland, australia: a bayesian spatio-temporal analysis, Epidemiology and Infection 145 (6) (2017) 1221–1230.

[8] Maracy, M. R., Mostafaei, S., Moghoofei, M., Mansourian, M., Impact of hiv risk factors on survival in iranian hiv-infected patients: A bayesian approach to retrospective cohort, Hiv and Aids Review 16 (2).

[9] Haario, H., Laine, M., Mira, A., Saksman, E., Dram: Efficient adaptive mcmc, Statistics and Computing 16 (4) (2006) 339–354. MR2297535

[10] Vrugt, J. A., Braak, C. J. F. T., Diks, C. G. H., Robinson, B. A., Hyman, J. M., Higdon, D., Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling, International Journal of Nonlinear Sciences and Numerical Simulation 10 (3) (2009) 273–290.

[11] O'Neill, P. D., A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo

methods, Mathematical Biosciences 180 (1) (2002) 103–114. MR1950750
[12] Mckinley, T. J., Ross, J. V., Deardon, R., Cook, A. R., Simulation-based bayesian inference for epidemic models, Computational Statistics and Data Analysis 71 (1) (2014) 434–447. MR3131981
[13] Demiris, N., O Neill, P. D., Bayesian inference for stochastic multitype epidemics in structured populations via random graphs, Journal of the Royal Statistical Society 67 (5) (2005) 731–745. MR2210690
[14] Bartlett, M. S., Deterministic and stochastic models for recurrent epidemics, University of California Press 4 (1956) 81–108. MR0084932
[15] Daniels, H. E., The distribution of the total size of an epidemic, Berkeley Symposium on Mathematical Statistics and Probability (1967) 281–293.
[16] Ball, F., Barbour, A. D., Poisson approximation for some epidemic models, Journal of Applied Probability 27 (3) (1990) 479–490. MR1067015
[17] Lefevre, C., Utev, S., Poisson approximation for the final state of a generalized epidemic process, Annals of Probability 23 (3) (1995) 1139–1162. MR1349165
[18] O Neill, P., Strong approximations for some open population epidemic models, Journal of Applied Probability 33 (2) (1996) 448–457. MR1385356
[19] Lefevre, C., Utev, S., Mixed poisson approximation in the collective epidemic model, Stochastic Processes and Their Applications 69 (2) (1997) 217–246. MR1472952
[20] Ball, F., Neal, P., Poisson approximations for epidemics with two levels of mixing, Annals of Probability 32 (1) (2004) 1168–1200. MR2044677
[21] Neal, P., The sis great circle epidemic model, Journal of Applied Probability 45 (2) (2008) 513–530. MR2426848
[22] Conesa, D., Martinez-Beneito, M. A., Amoros, R., Lopez-Quilez, A., Bayesian hierarchical poisson models with a hidden markov structure for the detection of influenza epidemic outbreaks, Statistical Methods in Medical Research 24 (2) (2015) 206. MR3336291
[23] Moneil, D. R., On the simple stochastic epidemic, Biometrika 59 (2) (1972) 494–497. MR0343396
[24] Kryscio, R. J., On the extended simple stochastic epidemic model, Biometrika 61 (1) (1974) 200–202.
[25] von Bahr, B., Martin-Lof, A., Threshold limit theorems for some epidemic processes, Advances in Applied Probability 12 (2) (1980) 319–349. MR0569431
[26] Fargette, D., Jeger, M., Fauquet, C., Fishpool, L. D. C., Analysis of temporal disease progress of african cassava mosaic virus, Phytopathology 84 (1) (1993) 91–98.
[27] Gatton, M. L., Cheng, Q., Modeling the development of acquired clinical immunity to plasmodium falciparum malaria, Infection and Immunity 72 (11) (2004) 6538.
[28] Konchom, S., Singhasivanon, P., Kaewkungwal, J., Chuprapawan, S., Thimasarn, K., Kidson, C., Yimsamran, S., Rojanawatsirivet, C., Early detection of malaria in an endemic area: model development, Southeast Asian Journal of Tropical Medicine and Public Health 37 (6) (2006) 1067–71.
[29] Kenah, E., Nonparametric survival analysis of epidemic data, Quantitative Biology (2012) 277–303.
[30] Shi, X., Ye, M., Curtis, G. P., Miller, G. L., Meyer, P. D., Kohler, M., Yabusaki, S., Wu, J., Assessment of parametric uncertainty for groundwater reactive transport modeling, Water Resources Research 50 (5) (2014) 4416–4439 (24).
[31] Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., Lloyd, A. L., Parameter estimation and uncertainty quantification for an epidemic model, Mathematical Biosciences and Engineering 9 (3) (2012) 553–576. MR2957535
[32] Box, G. E. P., Tiao, G. C., Bayesian Inference in Statistical Analysis, Wiley, 1992. MR0418321
[33] Haddad, K., Rahman, A., Stedinger, J. R., Regional flood frequency analysis using bayesian generalized least squares: a comparison between quantile and parameter regression techniques, Hydrological Processes 26 (7) (2011) 1008–1021.
[34] Larson, S. C., The shrinkage of the coefficient of multiple correlation, Journal of Educational Psychology 22 (1) (1931) 45–55.
[35] Breiman, L., Arcing classifiers, Annals of Statistics 26 (3) (1998) 801–824. MR1635406
[36] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, 1995, pp. 1137–1143.
[37] Verhulst, P. F., Notice sur la loi que la population suit dans son accroissement. correspondance mathematique et physique publiee par a quetelet, brussels, Quetelet 10 (10) 113–121.
[38] Richards, F. J., A flexible growth function for empirical use, Journal of Experimental Botany 10 (2) (1959) 290–301.
[39] Winsor, C. P., The gompertz curve as a growth curve, Proceedings of the National Academy of Sciences of the United States of America 18 (1) (1932) 1–7.
[40] Gelman, A. B., Carlin, J. B., Stern, H. S., Rubin, D. B., Bayesian data analysis, Wiley Interdisciplinary Reviews Cognitive Science 1 (5) (2014) 658–676. MR3235677
[41] Liu, W., Tang, S., Xiao, Y., Model selection and evaluation based on emerging infectious disease data sets including a/h1n1 and ebola, Comput Math Methods Med (1299) (2015) 1–14. MR3402199
[42] Choi, J., Harvey, J. W., Conklin, M. H., Use of multiparameter sensitivity analysis to determine relative importance of factors influencing natural attenuation of mining contaminants (1999) 1–7.

Zhu Meixia
School of Computer Science and Technology
Tianjin Polytechnic University
Tianjin, 300387
China
E-mail address: mxzhu@pku.edu.cn

Pei Yongzhen
School of Mathematical Sciences
Tianjin Polytechnic University
Tianjin, 300387
China
E-mail address: yongzhenpei@163.com

Ye Ming
School of Computer Science and Technology
Tianjin Polytechnic University
Tianjin, 300387
China
Department of Scientific Computing
Florida State University
Tallahassee, Florida
USA
E-mail address: mye@fsu.edu

Li Changguo
Department of Basic Science
Army Military Transportation University
Tianjin, 300161
China
E-mail address: bayeslimcmc@sina.com