Model Choice for Stochastic Epidemics in Households

C. H. Wen and P. H. O'Neill

***Abstract*—Although methods for parameter estimation for stochastic models of disease transmission are now well-established, the picture is much less clear for model assessment. We consider various approaches for model choice problems in the context of data on disease outbreaks collected at the level of individual households. The study provides practical values to identify feasible and numerical efficient model for given epidemic data.**

***Index Terms*—Model choice, stochastic epidemic models, reversible jump MCMC, Bayes factor.**

I. Introduction

Inferential framework for stochastic epidemic models have been well developed these years (for instance, [1] [2]), there are still few literature addressing the area of model choice. Especially, identifying an appropriate analysis tool for the specific data set is challenging. Model choice is concerned with the problem of distinguishing competing models. Various analysis tools have been designed and utilised in applications of model choice. Different criteria show their own bias on different statistics of model parameters. In this paper, we mainly study the applications of the different criteria and computational methods in the area of epidemic models. We aim to set up an effective scheme to identify a feasible criterion for specific epidemic data sets. Since MLE, MCMC and RJMCMC are associated with choice criteria in different models in practice, the technical problem arising from employing these computational methods will be investigated as well.

II. Model Choice Statistics

A. AIC

AIC initially proposed by Akaike [3] is used as a measure of information lost when a particular model is used in place of the unknown true model.

$$AIC = -2lnL_{max} + 2k, \tag{1}$$

where L_{max} is the maximum likelihood computed by the model and k is the number of parameters of the model. Given a data set, several competing models may be ranked according to their AIC and the best model is the one which minimises AIC.

B. BIC and DIC

To tackle the over-fitting problem, in which additional parameter increases the likelihood, BIC is introduced by Schwarz [4] to add a penalty term for the number of parameters in the model.

$$BIC=-2lnL_{max}+klnN \tag{2}$$

where N is the sample size.

As a combination of the logic from both Bayesian method and information theory, DIC is a hierarchical modelling generalisation of the AIC and BIC [5]. Let $D(\theta) = -2lnL + C$ where C is a constant dependent solely on data which will vanish from any derived quantity. We define $p_D = D(\theta) - D(\bar{\theta})$. The DIC is defined as

$$\text{DIC}=D(\theta) + p_D. \tag{3}$$

DIC is derived from the MCMC posterior samples. Therefore, the values of DIC are dependent on the priors employed during the MCMC simulations.

C. BF

Bayes Factor (BF) is another important tool which requires assessment of the sensitivity of the conclusion to the prior distribution. BF of model M1 over model M2 on the data space D is defined as

$$B_{12} = \frac{P(D|M1)}{P(D|M2)} \tag{4}$$

P(D|Mi) is the marginal likelihood for model i. If the models M1 and M2 are parametrised by vectors of parameters θ_1 and θ_2, we have

$$B_{12} = \frac{\int \pi(\theta_1|M1)P(D|\theta_1,M1)d\theta_1}{\int \pi(\theta_2|M2)P(D|\theta_2,M2)d\theta_2} \tag{5}$$

where $\pi(\theta_i|M_i), i = 1,2$ are the prior distributions.

III. Numerical Implementations

A. Independent Households

In the first case, we will implement the Longini-Koopman model [7] to study the applications of these selection tools to stochastic epidemic models.

We denote by B and Q respectively the probability of avoiding infection from the community and from the household. The community is composed of N households with four people. We have two outbreaks simulated with parameters B1, B2 and Q1, Q2. Given the epidemic data, we aim to choose one from the following two models:

- Model M1: B1 = B2 = B, Q1 = Q2 = Q
- Model M2: B1, B2, Q1, Q2

Then the likelihood function of model M1 is

$$L = \pi(B|M1)\pi(Q|M1)\prod_{i=0}^{4} p_{i4}(B,Q)^{n_{i4}} \tag{6}$$

where π(B|M1), π(Q|M1) are priors of B and Q, $p_{i4}(B,Q)$

Manuscript received May 12, 2012; revised June 18, 2012.

C. H. Wen is with university of Nottingham Ningbo, China (tel.: + (86)-574-88181407; fax: +(86)-574-88180175; e-mail: alan.wen@nottingham.edu.cn).

P. H. O'Neill is with university of Nottingham, Nottingham, United Kingdom.

is the probability of the household with i out of 4 infected and n_{i4} is the number of such households in N=100 households.

TABLE I: Simulation of Epidemic Data Using Different Parameter set B and Q, G1..G4 are group of Simulations and Ni are Simulated Epidemic data.

Group	E1	E2	E3	E4
G1 N_i	B:0.8, Q:0,8 41,21,16,14,8	B:0.45, Q: 0.75 4,9,16,30,41	B:0.75, Q:0.45 32,4,4,10,50	B:0.75, Q:0.75 31,17,18,17,16
G2 N_i	B:0.8, Q:0.5 41, 5,5,11,38	B:0.45, Q:0.75 4,9,16,30,41	B:0.75, Q:0.45 32,4,4,10,50	B:0.75, Q:0.75 31,18,17,18,16
G3 N_i	B:0.5, Q:0.8 6,13,21,31,19	B:0.45, Q:0.75 4,9,16,31,40	B:0.75, Q:0.45 32,4,4,10,50	B:0.45, Q:0.45 4,2,3,11,80
G4 N_i	B:0.3, Q:0.8 1,4,13,32,50	B:0.45, Q:0.75 4,9,16,30,41	B:0.75, Q:0.45 32,4,4,10,50	B:0.45, Q:0.45 4,2,3,11,80

TABLE I: Results of AIC, DIC, BF where G1…G4 refers to groups of results, and E:i-j refers to Simulation Data Ei and Ej in TABLE I.

Group	Outbreaks	AIC-M1	AIC-M2	DIC-M1	DIC-M2	BF: M1/M2
G1	E:1-2 E:1-3 E:1-4	640.5 595.8 616.6	573.9 540.6 615.8	640.5 595.7 616.6	574.0 540.5 615.8	$3.36e^{-14}$ $1.39e^{-13}$ 9.4
G2	E:1-2 E:1-3 E:1-4	585.3 502.0 601.7	536.3 502.9 578.1	585.2 501.9 601.7	536.2 502.9 578.0	$1.96e^{-10}$ 19.6 $9.6e^{-5}$
G3	E:1-2 E:1-3 E:1-4	577.8 595.2 508.6	577.9 544.6 454.7	577.8 595.2 508.6	577.8 544.6 451.4	8.2 $9.26e^{-11}$ $2.53e^{-12}$
G4	E:1-2 E:1-3 E:1-4	512.5 531.3 408.2	511.4 478.1 388.1	512.5 531.3 408.2	511.1 477.9 384.8	3.5 $1.73e^{-11}$ $4.82e^{-5}$

From the table, we reach consistent conclusions in model support by the statistics AIC, DIC and BF. However, the significance of support is more apparent by using BF than AIC and DIC.

1) *Prior Sensitivity*

We test the sensitivity of the BF values to the priors. We will see that the conclusion of the model choice with the tool of BF will be affected by the choice of the priors. We recall the outbreak E1 and E4 in different groups in the simulation above to see the BF value under different priors.

TABLE III: The BF values for the Implementation Results Using Different Priors for B and Q in Two Data Sets, Where we use Outbreak Data E1 and E4

Priors	M1:B	M1:Q	M2:B	M2:Q	BF
G1	B(4,1)	B(3,1)	B(4,1)	B(3,1)	3.1
G2	B(1,4)	B(1,3)	B(1,4)	B(1,3)	14.7
G3	B(10,40)	B(10,30)	B(10,40)	B(10,30)	$9.6e^{20}$
G4	B(0.4,0.1)	B(0.3,0.1)	B(0.4,0.1)	B(0.3,0.1)	73.3

We can see that the level of support varies quite dramatically with different priors although all results conclude that model M1 fits the data better than model M2.

In the following we use (B1=Q1=0.25, B2=Q2=0.3) , (B1=Q1=0.8, B2=Q2=0.75) to simulate two data sets. From the left figure, we can see that minimum logarithm of the BF values occurred in 0.28 with BF values 0.9363. It falls in the interval [0.24, 0.32] which is the curve of the logarithm of the BF intersecting the zero line. Intuitively, we believe the parameters are different if our knowledge assumes the true value is in the centre of the interval.

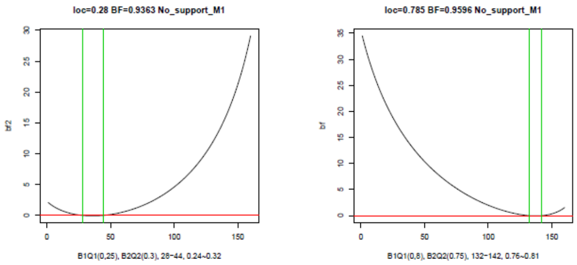

Fig. 1. Two Figures show the BF values will be influenced by the priors set for the computations.

In this case, support for M2 is within our expectation. We can see the minimum values falls nearly in the centre of the interval. A similar analysis applies to the right figure.

B. *Two-level Mixing*

Bayesian model choice methodology has been implemented to study the epidemic model with two level mixing. MCMC methods enable inference for such models, although the implementation details are usually non-standard [6]. We consider the two-level mixing epidemic model where the individuals are of only one type. This is one of the simplest and most basic kinds of heterogeneity. Epidemics in population mix at two levels: global and local. For each infective, there are two types of contacts to connect to the other individual: a global probability P_G , P_L for infecting each individual in the population and a local probability P_L for infecting each individual in his own household. An infective make contacts according to a Possion process during an infectious period *TI*, where *TI* are independently sampled from identical distribution *I*.

1) *Data and Likelihood*

The data are given as the form A = {a(n, i, f)} where a(n, i, f) denotes f frequencies of households which have initially n susceptible individuals and finally i of the n individuals get infected. The two basic parameters for this model are the infection rates λ_G and λ_L . By Bayes' theorem, the posterior density of the parameters satisfies

$$\pi\left(\lambda_{L,}\lambda_G \middle| A\right) \propto \pi(A|\lambda_L, \lambda_G) \cdot \pi(\lambda_L, \lambda_G) \tag{7}$$

Employing the technique of likelihood augmentation [8], we can derive the likelihood for λ_L and λ_G . We consider two models M1 and M2. Model M1 is the full model, with infection rate parameter λ_L and λ_G. Model M2 is identical but with only one parameter $\lambda = \lambda_L = \lambda_G/N$. Denote by P(M1) and P(M2) the two pre-assigned model priors. By employing the RJMCMC algorithm and counting the number of samples from each model, we have the following ratio

$$\frac{P(M1|A)}{P(M2|A)} = \frac{P(M1)\cdot P(A|M1)}{P(M2)\cdot P(A|M2)} \tag{8}$$

is nothing but B12, the Bayes factor of model M1 over model M2. Therefore the ratio computed in (8) via running RJMCMC is just the Bayes factor which evaluates the extent to which the given data support model M1 over model M2.

2) *Jump Proposal Mechanism*

We have the following jump proposal mechanism for the model switching in RJMCMC. If current model is M2, then we have the move $\hat{\lambda}_L = k\lambda + \mu,\ \hat{\lambda}_G = N \cdot \lambda$ and $\mu \sim N(0, \sigma^2)$. The Jacobian for the bijection mapping is N. Hence the jump probability is min{1, α} where

$$\alpha = \frac{\pi(\lambda_L,\lambda_G)\pi(\lambda_L,\lambda_G|A)r_1 N}{\pi(\lambda)\pi(\lambda,A)\pi_\mu\left(\hat{\lambda}_L - k\cdot\lambda\right)r_2}\,. \tag{9}$$

Similar numerics apply to the reverse move from M1 to M2.

3) *Numerical Results*

Denote by S={λ_L, λ_G, Z} the parameter sets and final size of each simulation. For presentation, one simulation S1={0.001, 0.35, 206} is given. The scale values K for S1 are 35. The standard deviation σ is 0.01.

TABLE IV: Top Table Refers to Simulated Small Household Data And Bottom Table Refers to the BF Values Given the Data Set Which Are Influenced by Priors of the Parameters

House Size	**1**	**2**	**3**	**4**	**5**	**Total**
Frequency	30	50	30	20	10	140
Total	30	100	90	80	50	350

K	V_L	V_G	α	β	B_{12}
35	0.1	0.01	1.39	4.07	660/9340
35	1.0	1.0	1.79	205.72	1879/8181
35	10.0	1.0	1.39	407.25	5567/4433

where $V_L, V_G\ \alpha, \beta$, are prior parameters. From the results listed above, we can see that the ratio derived from RJMCMC is heavily influenced by the priors given to the model parameters.

References

[1] P. D. O'Neil and G. O. Roberts, "Bayesian inference for partially observed stochastic epidemics," *J. R. Statist. Soc. A*, vol. 162, no. 1, pp. 121-129, 1999.

[2] P. D. O'Neill and C. H. Wen, "Modelling and inference for epidemic models featuring non-linear infection pressure," *Mathematical Biosciences*, 2012.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716-723, 1974.

[4] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.

[5] D. J. Spiegelhalter, D. G Best, B. P. Carlin, and A. Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society B*, vol. 64, no. 4, pp. 583-639, 2002.

[6] P. D. O'Neill, "Bayesian inference for stochastic multitype epidemics in structured populations using sample data," *Biostatistics*, vol. 10, no. 4, pp. 779-791, 2009.

[7] I. M. Longini and J. S. Koopman, "Household and community transmission parameters from final distributions of infections in households," *Biometrics*, vol. 38, pp. 115-126, 1982.

[8] N. Demiris and P. D. O'Neill, "Bayesian inference for stochastic multitype epidemics in structured populations via random graphs," *Journal of the Royal Statistical Society*, Series B 67, pp. 731-746, 2005.