International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 6, Issue2, pp: (1-9), Month: October 2018 - March 2019, Available at: www.researchpublish.com

MODELLING OF CEREBRAL SPINAL MENINGITIS (CSM) OUTBREAK USING STOCHASTIC EPIDEMIC MODELS WITH BAYESIAN PRIORS

ABDUL BADI SAYIBU

University Of Ghana, Legon, Ghana

Abstract: **Data analysis is undoubtedly a major part in any general epidemic preparedness and control. Several statistical methods have been developed over the years for better and accurate analysis of epidemic data. Key among such methods is the use of stochastic epidemic models. This study used Bayesian technique to do inference for stochastic epidemics given data from CSM outbreak. The data for the analysis was a 2010 CSM outbreak data from Jirapa township in Ghana with a population of 6139 people and an ultimately infected number of 354 people. Conjugate and uninformative prior selections methods were considered for the study of which the former was seen to present better results based on the diagnosis analysis of results from both methods. In estimating the summaries of these parameters from the joint conditional distribution, the study adopted MCMC algorithms which was implemented in R using a combination of packages. In presenting the results, the study considered 2 parameters out of the several that can be estimated with the model. The results of the sampling of the exposure period E shows that the outbreak population has a very high exposure period with E having a median of 8.8965 days. This is indeed very high as compared to the WHO standard of between 2 and 10 days. Also, the rate of infection β had a median of 0.7334. The two parameter estimates are some of the factors that lead to the ultimate extinction of the outbreak before it could spread to all the entire township.**

Keywords: **Epidemics, CSM, stochastic, MCMC algorithm, random graph, Bayesian theorem, social network.**

I. INTRODUCTION

Methods for the modelling of infectious disease propagation have been a challenging field for many researchers for decades, partly due the way data collection is conducted during any major epidemic outbreak, and partly due to the complex nature of population structures and most especially that of human populations.

O'Neill (2002) described the analysis of disease propagation as a non-standard problem of which so many different methods have been proposed and used in some cases. The desire for accurate analysis of epidemic data have led to so many different approaches being proposed.

The use of Bayesian method for the analysis of epidemic outbreak data using stochastic epidemic models have gained so much interest among researchers in recent years. Epidemic outbreak by its nature is a stochastic phenomenon in which the disease can either get extinct after initial emergence or will ultimately infect many individuals in a given population. Secondly, attention can be due to the flexible nature of the method. The method assumes randomization for parameters which can be estimated at any small time $t > 0$ given some data. This makes the Bayes' method much more consistent with the behavior of epidemic outbreak in a population, of which the parameters of the epidemic and the population can change at any given time.

Even though a lot of work have been done on the use of Bayes' theorem in the analysis of epidemic data using stochastic models, much of the work largely remains at the methodological development stages with little or no attempt at the use of these methods to do inference using real epidemic outbreak data.

O'Neil (2002) did some work on a model that was initially described by Andersson (1999). The model was later extended in the work of Hunter *et al.*, (2010) with inference on Measles outbreak data from Hagelloch, a community in Germany in the 1800s. The work done on the model have largely been the development of the methodology needed to estimate both the epidemic parameters and the those of the population structures. The model assumes a random population structure which is described in the model using a Bernoulli random graph.

This study therefore performed inference on the epidemic model as described in O'Neill (2002) using Cerebrospinal Meningitis (CSM) outbreak data which occurred in Ghana. The data will be a good one for inference using such model since the transmission process of CSM in Ghana is largely seen to be by human contact, which is one of the basic assumptions of the model among others.

II. METHODOLOGY

Analysis of epidemic data have traditionally been very challenging especially giving the population structures of an infected population. Hunter *et al.*, (2010) noted in their work that the main challenge is how to difference the impact of the population structure parameter and epidemic parameters on the spread of the epidemic. O'Neill (2002) also noted the difficult of striking a balance between a too simplistic model that might not be attached with any seriousness to run a meaningful inference, and a far too complex model that ends up given too much information that it is eventually very difficult to understand.

The basic assumption of this model is that the epidemic outbreak is over. This implies the number of observe removals is the same as those infected.

By the compartmental modelling method, the entire population can be divided into four groups of Susceptible, Exposed, Infected and Remove, generating a model called the SEIR model.

If $R = (R_1, R_2, \dots, R_N)$, $I = (I_1, I_2, \dots, I_N)$ and $E = (E_1, E_2, \dots, E_N)$ represents the random variables associated with the Exposure, Infected and Removal times of individuals in the population respectively. Then, the epidemic data of the entire population will be of the form:

$$EIR = [(E_1, I_1, R_1), (E_1, I_2, R_2), \dots, (E_N, I_N, R_N)] \quad (1)$$

To deal with the challenge of modelling the population structures, the model makes some convenient assumptions which include a uniform probability of contact between any two individuals and the independence of the existence of a contact between any two individuals. The contact process can therefore be assumed to follow an Exponential-Family Random Graph Model (ERGM) given by the relation:

$$G(N, p) = \frac{exp\{\sum_{t=0}^{k} p_t \sum_{\{u,v\}} G_{\{u,v\}} X_{\{u,v\},t}\}}{k(p)} \quad (2)$$

Where $G_{\{u,v\}}$ indicates the presence of a contact between u and v and X is a matrix of $\binom{N}{2} \times k$ of the dyadic covariates.

With the intension of limiting the model to the exponential family of distribution, the various epidemic parameters are identified as follows:

Given that the infectious make contact with susceptible individual at an exponential rate with mean $\frac{1}{\beta}$, then $\beta > 0$ is the rate of infection.

The study also used the gamma distribution to describe the random variables associated with the exposed and infectious states. That is, the time an individual spends at each state. If E is the random variable associated with the exposure time, then $E \sim gamma(\theta_E, k_E)$. Equivalently, if I is the random variable associated with the infection period, then $I \sim gamma(\theta_I, k_I)$. These variables are included in the model later as parameters to be estimated.

Two kinds of prior distributions selections will be considered. The conjugate prior and the non-informative prior for both the network and model parameters. Specifically, beta distribution is used for the network parameter p which is conjugate for the distribution of p; the gamma distribution is conjugate for the Model parameters β which has an exponential distribution. Also, inverse gamma distribution is assigned for the model parameters θ_I and θ_E since they both are parameters of a gamma distribution.

However, all the above parameters will also be assigned uninformative uniform priors during the analysis to compare which of the two methods improves the model.

Bayesian inference method was then used to estimate the parameters of the model. This was done by assigning prior distributions for the various parameters and together with the available data, the joint posterior can be estimated using the relation of the Bayes theorem given as:

$$f(\theta \mid data) = \frac{f(data \mid \theta) f(\theta)}{f(data)} \tag{3}$$

Where θ is the set of parameters for the joint posterior distribution.

Based on the Bayes theorem as given above, the joint posterior distribution is therefore given by;

$$\pi(\Omega \mid X) \propto L(X \mid \Omega)\pi(\beta)\pi(\eta|p)\pi(\delta|G)\pi(p)\pi(\theta_I)(k_I)\pi(\theta_E)\pi(k_E)\pi(I)\pi(E)\pi(k) \tag{4}$$

This includes all the possible parameters of the model including the realizations of the network parameters G and p given as δ and η respectively. $\Omega = \{\beta, \eta, \delta, \theta_I, k_I, \theta_E, k_E, E_k\}$ represents the set of all parameters.

An MCMC algorithm was then used to sample these model parameters, and summary statistics generated from the sample values of each parameter. This is particularly necessary to avoid the complex calculations that are involved in estimating the likelihoods of various parameters as can be seen in the Bayesian paradigm.

Finally, the Model was implemented using a combination of R packages including epinet, a package that is built base on the proposed algorithm choice for this study being the leading package. The performance of the model was tested by doing a diagnostic analysis for the MCMC output. This involved checking for various model performance indicators such as mixing and convergence. For more detail on how the R package epinet works see Christ and David (2014).

III. RESULTS AND DISCUSSIONS

A preliminary analysis done to achieve two things, first to subject the epidemic data to consistency checks to ensure that there are no outliers which will end up affecting the accuracy of the epidemic model. On the second part of the preliminary analysis, the infection and removal time of each of the 354 ultimately infected individuals are estimated based on the case reporting date and date of onset of symptoms. The unit of measurement of time in the case of the CSM data is days.

Table 1: Summary of various variable in the CSM data set

Var. name	obs.	mean	median	s.d.	min.	max.
id	354	177.5	177.5	102.34	1	354
age	354	23.18	18	19.3	0.17	81
sex	354	1.489	1	0.501	1	2
fever	344	1.96	2	0.195	1	2
headache	354	1.986	2	0.118	1	2
stiffneck	354	1.986	2	0.118	1	2
lp	354	1.757	2	0.436	1	2
reported-date	350	6-10-2010			2-7-2010	8-5-2010
onset-date	353	6-7-2010			2-5-2010	8-5-2010
outcome	354	1.096	1	0.295	1	2

The data as received from the Ghana Health Service is first subjected to quality check to ensure all duplications and missing values are corrected. The variables of interest that will enable us to do this is the identity or id of the individuals.

Taking a closely at the summary of the identity of each of the 354 individuals. The summary has a minimum value of one and a maximum value of 354 which is accurate. Furthermore, we check the mean of the variable by obtaining mean (1:354) which returns a value of 177.5, consistent with what the summary produced.

All the variables relating to symptoms are coded as factors with a yes or no responds for each. The case definition is the appearance of any of the symptoms as shown in table 4.1

Since the stochastic epidemic models are inferred using time data, the original CSM data will be used in doing that in this section. The original CSM data has the date each of the infected individuals first showed symptoms of CSM and the date that they reported to a health facility.

To estimate the removal and infection times, it is important to first make two assumptions with expert advice from the surveillance unit of the Ghana Health Service. The study assumes that infection started a day before symptoms first emerged and removal time began 8 days after a person reports to a health facility, after which the person should have become well or died.

After estimating the removal and infection times of the ultimately infected individuals, the information is put back into the population data to obtain an $N \times 5$ matrix, where N is the total number of individuals in the affected population. The five columns represent respectively the individual ID, the parent node most likely responsible for their infection, the exposure, infection and removal times. All missing data and data for uninfected individuals in the population are coded with NA.

Below is table 2 showing a summary of the final data as used for the inference.

Table 2: Summary of infection and removal times

	NodeID	Parent	Etime	Itime	Rtime
Min.	1	NA	NA	3.00	2.0
1st Qu.	1536	NA	NA	9.00	135.0
Median	3070	NA	NA	10.00	149.0
Mean	3070	NaN	NaN	10.51	41.1
3rd Qu.	4604	NA	NA	11.00	160.0
Max.	6139	NA	NA	32.00	207.0
		NA's: 6139	NA's: 6139	NA's: 5789	NA's: 5789

From Table 2, it can be observed that the data obtained from the Ghana Health Service was only enough to construct the infection and removal time of 350 out of the 354 recorded cases during the entire outbreak period. This is due to missing time records for 4 infected individuals.

For the purpose of including the population parameter in the epidemic model, we considered every member of the population as a data point regardless whether their information is available or not. To this regard, the NodeID variable ranges from 1 to 6139 with each individual in the population assigned a unique ID.

Exposure time of all the members of the population whether infected or not is not known and so is coded NA. This is equally true for the variable Parent. The variable Parent is supposed to identify the most likely individual that is responsible for infecting a particular Node. Again, this is coded NA for all uninfected individuals and infected individual for which such data is not available.

After satisfying all basic assumptions of the model in terms of obtaining the required data format, the actual implementation of the model was conducted.

The algorithm was implemented and diagnosed using a combination of packages in R. The main packages being epinet package and coda package. The algorithm was run for 1000000 times and thinning at an interval of 1000 iteration to produce a minimum of 10000 samples for each parameter.

The algorithm was first run with uninformative Uniform prior for all model parameters, then with the respective conjugate priors as specified in early in the study. For both cases, the transmission tree is maintained as having a Uniform prior.

Specifically, we choose distributions as follows:

$\pi(\beta|.) \sim U(0,2)$, $\pi(p|.) \sim U(0,1)$, $\pi(\theta_E|.) \sim U(2,10)$, $\pi(\theta_I|.) \sim U(0.2,0.5)$, $\pi(K_I|.) \sim U(12,20)$ and $\pi(K_E|.) \sim U(2,10)$, all for the uninformative Uniform prior.

For the conjugate prior, we choose the prior as follows;

$\pi(\beta|.) \sim Gamma(2,2)$,

$\pi(p|.) \sim beta(1,1)$,

$\pi(\theta_E|.) \sim IGamma(2.5,1)$,

$\pi(\theta_I|.) \sim IGamma(0.2,0.5)$, $\pi(K_I|.) \sim IGamma(8,20)$ and $\pi(K_E|.) \sim IGamma(2.5,1)$.

The parameter specifications are done randomly but to accommodate some constrains that have been imposed by some distributions. For instance, the beta distribution can take values 0 to 1. Similar constrains can be said of invers Gamma, Gamma and Uniforms distributions respectively.

In diagnosing the output of the MCMC algorithm, we take a first look at the first scenario in which noninformative uniform prior was used for all epidemic related parameters. We take a look at the summary statistics of selected parameters from the joint posterior distribution. Of particular interest is the parameters relating to the time of exposure E and that of the rate of infection β. In addition to the exposure time and the rate of infection parameters, we also sampled 2 parameters relating prior distribution of the exposure time variable E. That is K_E and θ_E. From the summaries in table 4.3 below, the exposure times of ultimately infected individuals have a mean and median of 6.9586 and 6.9732 respectively. This is consistent with available literature and regulations on CSM which puts the incubation period between 2 to 10 days.

The rate of infection β also appears to be within a reasonable range, having a mean of 0.97589 and a median of 0.98589. The rate of infection β is justifiably small due to the fact that we had just a total of 350 ultimately infected individuals from a population of 6139 individuals.

Table 3: Posterior summary of selected parameters from the uniform prior

	β	K_E	θ_E	E
Min.	0.02589	2.237	3.702	0.5544
1st Qu.	0.88844	3.366	5.237	7.5654
Median	0.98341	3.639	5.674	6.9732
Mean	0.97589	3.655	5.735	6.9586

For further diagnoses of the model in terms of mixing and convergence, we take a close look at the acceptance rate and effective sample size of the selected parameters. The acceptance rate of all four parameters ranges from 0.2980 to 0.2707 indicating a good mixing and convergence of the algorithm. As discussed earlier, an acceptance rate of between 25% to 75% indicates a good mixing and an ultimately converging algorithm.

Furthermore, an examination of the effectively sample size indicates a good mixing and converging algorithm. The effective sample size is the sample size of each parameter taking into consideration autocorrelation. Generally, an effective sample of 200 or more is enough to obtain the needed information about a given parameter. An algorithm therefore needs to run long enough to obtain such a sample. In the case of this study's algorithm, it was run 1000000 times, enabling a sample of nearly 10000 or more for each parameter. One thing to notice from the output of the acceptance and effective sample is that all the three parameters relating to the Exposure times have the same values. This is understandable since there all are geared toward explaining the same epidemic phenomenon. Table 4 below shows the effective sample size and the acceptance rate of selected parameters using uniform priors.

Table 4: Effective sample size and acceptance rate of parameters with Uniform prior

	β	K_E	θ_E	E
Acceptance rate	0.2707	0.2985	0.2985	0.2985
Effective sample size	10107.00	10980.5151	10980.5151	10980.5151

The posterior summaries of the same parameters using their respective conjugate priors produced a rather slightly different result compared to those sampled using uniform prior. The exposure period E have mean of 8.8893 and median of 8.8965, both significantly higher than those produced with the uniform prior.

On the other hand, the rate of infection β have a mean of 0.72964 and median 0.73340, both significantly lower than those produced using the uniform noninformative prior. In both cases however, the results are within existing ranges that existing knowledge and literature on CSM supports.

The issue of which prior selection technique makes the model better will be explored further in the other diagnosis methods of the MCMC results. Below is the table showing summaries of the parameters from the conjugate prior distributions.

Table 5: Summary of selected parameters from the conjugate priors

	β	K_E	θ_E	E
Min.	0.02218	5.000	2.000	0.2568
1st Qu.	0.65438	6.412	2.955	9.5125
Median	0.73340	6.555	2.977	8.8965
Mean	0.72964	6.560	2.968	8.8893

Like the output obtained with a uniform prior, the acceptance rate for the output with conjugate priors ranges between 0.3180 and 0.3307 for all selected parameters. Also, the effective sample size ranges from 10180.3220 to 10307 across all four parameters.

These figures show a slight improvement compared to those obtained using uniform prior for all parameters. This is particularly visible in the effective sample size with which there is a generally slight improvement compared to those obtained from the uniform prior output. Below is table 6 showing the output for the acceptance rate and effective sample size for the four selected parameters.

Table 6: Effective sample size and acceptance rate of parameters with conjugate priors

	β	K_E	θ_E	E
Acceptance rate	0.3307	0.3180	0.3180	0.3180
Effective sample size	10307.00	10180.3220	10180.3220	10180.3220

Finally, the marginal posterior density of β and K_E as sampled from the joint posterior distribution is considered.

First take a look at the posterior density of β and K_E. The algorithm was able to recover the initial values of zero for each of the two parameters. The posterior density also shows rate of infection going as high as 2.0 at some point of the epidemic making it the maximum of value for β. Below are the marginal posterior densities for β and K_E respectively.

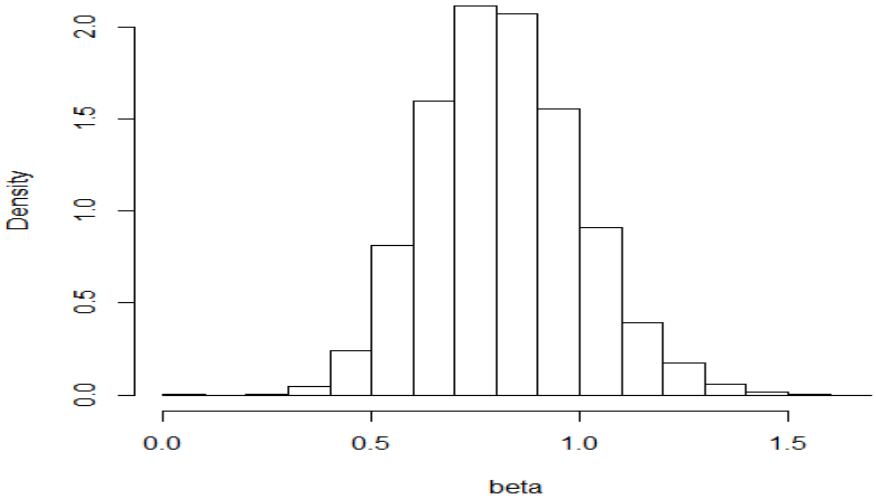

Fig.1: Marginal posterior densities for β using conjugate prior

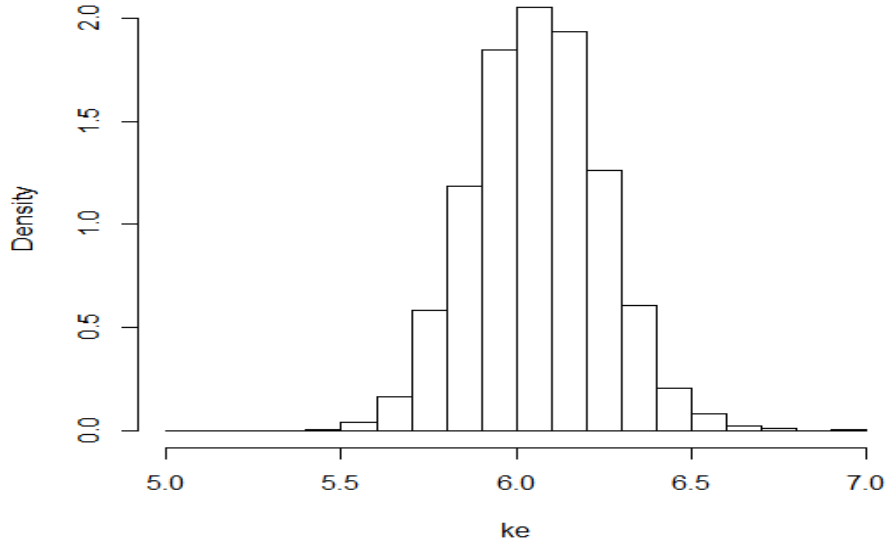

Fig.2: Marginal posterior densities for K_E using conjugate priors

One indicator of interest of the algorithm is the autocorrelation functions of the samples of respective parameters. Again, the interest is on the rate of infection β. The study also considered the parameter θ_E. By autocorrelation, looks at the correlation between every sample and the sample n iterations before it. In a well-mixed and converged algorithm, the autocorrelation will get smaller with increasing iterations. In other words, when the algorithm runs long enough it should be able to produce independent samples for each of the parameters of the model. Higher autocorrelation at higher iterations is an indication of a slow mixing algorithm.

The autocorrelation of β and θ_E as shown in Fig. 3 and Fig. 4 indicates one that approaches zero as the number of iterations increases. In both cases, autocorrelation is zero by the expected minimum sample of 10000 indicating that the algorithm was producing independent samples for each parameter after that. Below are Fig. 3 and Figure 4 showing the autocorrelation functions of the parameters β and θ_E respectively.

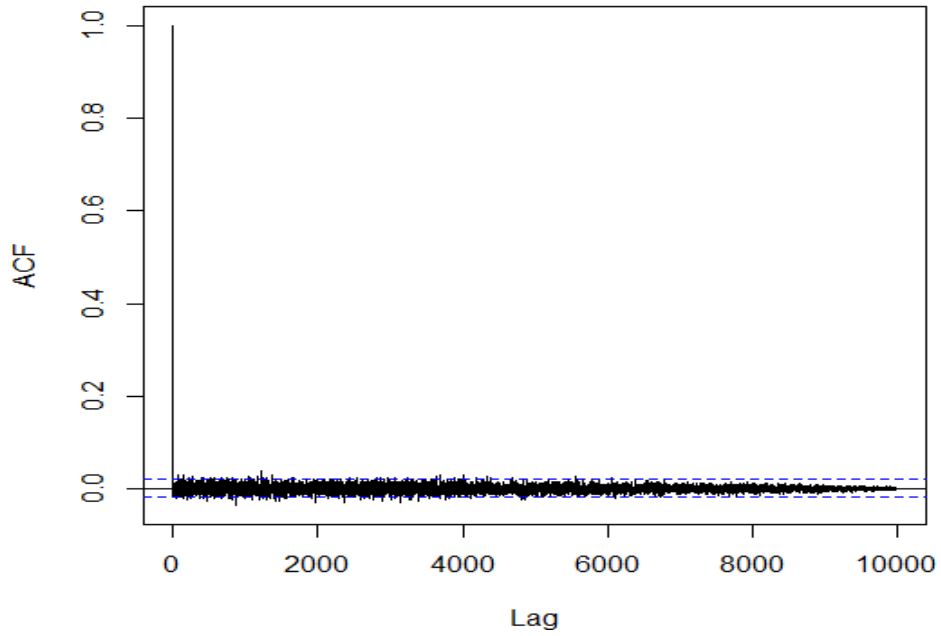

Fig.3: Autocorrelation function for β using conjugate priors

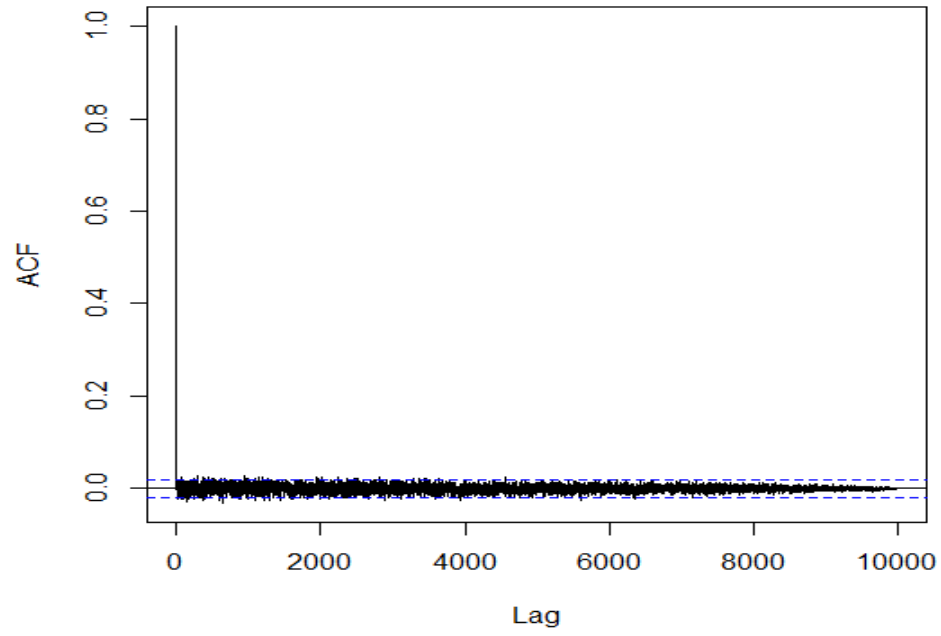

Fig. 4: Autocorrelation function for θ_E using conjugate priors

IV. CONCLUSION

The study conclusion starts with the model selection. Two set of Bayesian prior selection method were considered for the study, Uniform priors and conjugate priors. Based on the diagnoses presented in chapter 4, the conjugate prior models are selected over the Uniform prior. Although both prior selection methods produce acceptance rate that is between the acceptable range of 25-75, the conjugate prior produced larger effective sample size as compared to the Uniform prior. For all parameters, the conjugate prior produced an acceptance rate from between 0.3180 and 0.3307 whilst the Uniform produced an acceptance rate from between 0.2807 and 0.2985. The effective sample size for conjugate priors ranges from 10180-10307 whilst that of the Uniform priors ranges from 10107-10980.5151.

The analysis therefore revealed a very high exposure period for the population under consideration with E having a mean of 8.8893 and a median of 8.8965. This means that the CSM virus takes an average of 8.8893 days to incubate in individuals after an initial contact.

ACKNOWLEGEMENTS

My very first sincere appreciations go to the Almighty God for seeing me through a whole year of research, given me the strength in terms of health and the wisdom to be able to put together all the information contained in this research.

I cannot also forget to say a big thank you to Dr. Louis Asiedu for all the time you dedicated out of your very busy schedule to respond to my queries throughout the research.

To all my friends in the statistics department, most especially John, Nafisah, Ama and Christophilia, I really appreciate your support.

REFERENCES

[1] Ahmed, E. A. (2015). Estimation of some lifetime parameters of generalized Gompertz distribution under progressively type-II censored data: Applied Mathematical Modelling

[2] Anita, S. (2005). An optimal parallel algorithm to construct a tree 3-spanner on interval graphs. International Journal of Computer Mathematics

[3] Arkady, S. and Alexander, K. (2017). Foundations of Bayesian Analysis: Wiley

[4] Bernardo, J.M. and Smith, A.F.M. (1994). Bayesian Theory. Section 5.2

[5] Britton, T. (2002). Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. Scandinavian Journal of Statistics

[6] Centers for Disease Control and Prevention (2012). Principles of Epidemiology, Third Edition. Atlanta, Georgia

[7] Chris, G. (2010). Bayesian Inference for Contact Networks Given Epidemic Data: Network inference using epidemic data. Scandinavian Journal of Statistics

[8] Chris, G. (2012) A Network-based Analysis of the 1861 Hagelloch Measles Data. Biometrics

[9] Chris, G., David, W. and David, R. H. (2010). Bayesian Inference for Contact Networks Given Epidemic Data: technical reports and preprints

[10] Christ, G. and David, W. (2014). An R Package to Analyze Epidemics Spread Across Contact Networks

[11] Cranmer, S. J. Desmarais, B. A. (2011). Inferential Network Analysis with Exponential Random Graph Models. Political Analysis. 19 (1): 66–86. doi:10.1093/pan/mpq037

[12] De Oliveira, M. L., Mallo, D. and Posada, D. (2014). A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction: Systematic Biology

[13] Del Moral, P. and Lawrence, M. M. (2015). Sequential Monte Carlo with Highly Informative Observations. SIAM/ASA Journal on Uncertainty Quantification

[14] Demetris, L. (2009) Trans dimensional Sampling Algorithms for Bayesian Variable Selection in Classification Problems with Many More Variables Than Observations. Journal of Computational and Graphical Statistics

[15] Freeman, L. (2004). The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press. ISBN 1-59457-714-5

[16] Geoff, P. G. (1997). The Natural History of Syphilis: Sexually Transmitted Diseases

[17] Gerson, M. (2015). The next epidemic. The Greenville News. South Carolina, US.

[18] International Federation of Red Cross and the Crescent Societies (2012), DREF final report: Ghana Meningitis outbreak

[19] Kofi, I., Alexis, N. and Chris, F.O. (2011). Maternal and neonatal survival and mortality in the Upper West Region of Ghana. International Journal of Gynecology & Obstetrics

[20] Marina, Y. S. and Nadezda, A. K. (2011). Stochastic perspective of industrial distribution network processes: Industrial Marketing Management

[21] Martin, P. M., Martin-Granel, E. (2006). 2,500-year evolution of the term epidemic. Emerging Infect. Dis. **12** (6): 976–80. PMC 3373038. PMID 16707055.

[22] Nikolaos, D. (2004). Bayesian Inference for Stochastic Epidemic Models using Markov chain Monte Carlo Methods

[23] Nikolaos, D. (2005). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. Journal of the Royal Statistical Society Series B (Statistical Methodology)

[24] O'Neill, D. P., and Britton, T. (2002). Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure

[25] O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods: Mathematical Biosciences

[26] O'Neill, P. D., Balding D. J., Becker N. G., Eerola M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods. J. Roy. Statist. Soc. Ser. C 49, 517–542

[27] Reinert, G. (2009), Statistical Inference for Networks: Graduate Lectures

[28] Sajja, "Front Matter", Chapman & Hall/CRC (2012). Data Mining and Knowledge Discovery Series

[29] Scott, J. P. (2000). Social Network Analysis: A Handbook (2nd edition). Thousand Oaks, CA: Sage Publications

[30] Scott, M. L. (2007). Basics of Bayesian Statistics: Statistics for Social and Behavioral Sciences

[31] The New Palgrave Dictionary of Economics (2008). The New Palgrave: Dictionary of Economics, Volume 5 Lardner — network goods (theory)

[32] Wasserman, S. and Faust, K. (1994). Social Network Analysis in the Social and Behavioral Sciences. Social Network Analysis: Methods and Applications. Cambridge University Press. pp. 1–27. ISBN 9780521387071

[33] World Health Organization (2012). Controlling the global obesity epidemic

[34] Yildirim, I. (2012). Bayesian Inference: Metropolis-Hastings Sampling