Estimation of under-reporting in epidemics using approximations

Estimation of under-reporting in epidemics using approximations

Kokouvi Gamado[1,*], George Streftaris [2], Stan Zachary [2]

[1] Biomathematics and Statistics Scotland

[2] Maxwell Institute for Mathematical Sciences, Heriot-Watt University

[*] Kokouvi.Gamado@bioss.ac.uk, ☎ 01316504902

August 1, 2016

Abstract

Under-reporting in epidemics, when it is ignored, leads to under-estimation of the infection rate and therefore of the reproduction number. In the case of stochastic models with temporal data, a usual approach for dealing with such issues is to apply data augmentation techniques through Bayesian methodology. Departing from earlier literature approaches implemented using reversible jump Markov chain Monte Carlo (RJMCMC) techniques, we make use of approximations to obtain faster estimation with simple MCMC. Comparisons among the methods developed here, and with the RJMCMC approach, are carried out and highlight that approximation-based methodology offers useful alternative inference tools for large epidemics, with a good trade-off between time cost and accuracy.

Key words– Under-reporting; Approximations; Markov Chain Monte Carlo; Reversible jump; Stochastic SIR model; Final size distribution

1 Introduction

The generalised stochastic epidemic has been studied in different ways for inference purposes. Developments in such studies include the Bayesian approach using Markov Chain Monte Carlo (MCMC) techniques to provide inference about the model parameters and the basic reproduction number (Heesterbeek and Dietz, 1996), particularly in the case of partial observations (Gibson and Renshaw, 1998; O'Neill and Becker, 2001; Streftaris and Gibson, 2004a). Such studies do not account for possible under-reporting of infected cases, i.e. when the data do not reveal the true number of infections that actually occur in the

population. In other words, there could have been a number of infected cases that are not reported even after the epidemic has ceased. Epidemiological and socio-economic factors are among the main reasons of under-reporting (Gamado *et al.*, 2014).

Under-reporting is very important in inferential problems and has been considered in many situations. Clarkson and Fine (1985) used time series data to estimate the efficiency of measles and pertussis notification reporting by comparing annual number of births and notifications with modification including detailed age-specific data (see Bjornstad *et al.*, 2002). For the novel influenza A (H1N1), early findings methodology developed by Fraser *et al.* (2009) considered the daily cases observed with potential under-reporting rate, to estimate the reproduction number at each given day. Such approach is mainly based on the observed final size which is binomially distributed, conditional on the actual final size, with parameters given by the true unobserved number of cases and the rate of reporting. Hens *et al.* (2011) considered a non-parametric approach where, by looking at the daily number of cases, they pre-smoothed the cumulative number of cases to infer the missing update. White and Pagano (2010) investigated the impact of under-reporting on estimates of both R_0 and the serial interval using likelihood-based approach methodology and daily cases of the epidemic. More recent work by Dorigatti *et al.* (2012) couples a deterministic mathematical model with a statistical description of the reporting process, with application to surveillance data collected in Italy.

Although the issue of under-reporting in epidemics has been widely studied, most of the emphasis in literature work has relied on aggregate data (e.g. epidemic final size) combined with empirically calibrated model parameters. There has been limited attention on integrated efforts where full inference on both the epidemiological characteristics and the under-reporting rate is considered. Here we present such an approach based on Bayesian methodology and using detailed temporal data. In what follows, we consider the Markovian SIR (Susceptible, – Infected, – Removed) epidemic in which removals are reported with probability p independently for all individuals in a population. SIR models often represent a simplification of realistic circumstances; however, they are still particularly relevant and useful, especially for methodology development (Xu *et al.*, 2016; Neal and Huang, 2015; Knock and O'Neill, 2014) when new methodological features necessitate the consideration of computationally intensive assessment through repetitive simulations. Various factors that can influence the reporting process through time or within social networks are not considered here, and therefore we assume constant reporting probability, p. The impact of such factors in the epidemic process and related parameter estimation has been explored in earlier work (Gamado *et al.*, 2014). We assume that reporting coincides

with removal, i.e each removal time is independently reported with probability p. The methodology developed throughout this paper is applicable to the generalised stochastic epidemic model where different distributions can be used for the infectious period.

A usual approach to estimation is to derive the full likelihood of the model, which requires imputing all unreported infection and removal times, together with all unobserved infection times for reported removals in the case of partial observation. Such an approach was explored in Gamado *et al.* (2014) with the use of Reversible Jump MCMC (RJMCMC). However, one difficulty associated with this approach is the impracticality of the RJMCMC even for moderately large populations. MCMC methods have been applied to large population sizes (Jewell *et al.*, 2008; Chis-Ster and Ferguson, 2007). Here, due to under-reporting in the epidemic, involving unknown event times coupled with a large number of individuals that might or might not be infective, RJMCMC becomes inefficient, as it requires the imputation of a very large number of variables. In cases where early inference on epidemics is required, time-efficient methods are necessary, as any intervention and mitigation measures will need to rely on timely and reliable estimates of the key characteristics of the outbreak dynamics, including the extent of under-reporting of infectious cases. The example of the recent Ebola epidemic is very edifying. According to WHO epidemiologists working in Guinea, the true scale of the outbreak was being underestimated by officially reported data, though the magnitude of under-reporting could not be accurately measured (World Health Organization, 2015). The framework developed in this manuscript can be applied in similar situations of large-scale populations, when the local spread of the disease is confined within a well defined area.

We are thus motivated in this paper to consider approximations which should work efficiently and accurately for large population size. Such approximations will help by allowing inference for parameters of interest, based on reported cases only, using MCMC methods which avoid the computationally intensive changes of dimension involved in the RJMCMC methods Gamado *et al.* (2014).

The remainder of this paper is organised as follows. In Section 2 we first present the model, state clearly the various approximations to be used. An approximate likelihood is then derived as the basis for implementing Bayesian inference to estimate the model parameters in Section 3. An inferential technique using the approximate likelihood is described in Section 3.1, and a suitable adjustment is made in Section 3.2 in order to allow for uncertainty about p. We also consider an alternative estimation approach employing Gibbs-like steps to make inference for the model parameters in Section 3.3. In Section 4 we apply the different methods to data and compare with full RJMCMC estimation.

Simulation studies are provided in Section 5 before adding discussion in Section 6.

2 Model and approximations

2.1 Model description

In a closed population of size N, we consider the Markovian SIR model in which removals are reported with probability p, independently over all individuals. Under full reporting of removals, the dynamic transition probabilities between states are given by

$$\Pr\{S(t+dt) = S(t) - 1,\ I(t+dt) = I(t) + 1\} = \beta S(t) I(t)\, dt + o(\, dt) \quad (1)$$

$$\Pr\{I(t+dt) = I(t) - 1,\ R(t+dt) = R(t) + 1\} = \gamma I(t)\, dt + o(\, dt), \quad (2)$$

where the parameters β and γ are respectively the contact and removal rates; $I(t)$ and $S(t)$ represent the number of infectious and susceptible individuals respectively at time t. In a formulation based on individual cases (Neal and Roberts, 2005), the process in (2) is equivalent to assuming that the infectious period is an exponential Exp(γ) random variable.

For this Markovian model it clearly does not matter whether we are able to pair the infection and removal times, but for specific distributions of the infectious lifetime except the exponential, this information would improve inference by allowing, for instance, non-centered parameterisations (Papaspiliopoulos *et al.*, 2003; Neal and Roberts, 2005).

On top of the physical progression of the epidemic we consider a reporting process, aiming to develop inferential analyses based solely on reported cases, rather than on all possible cases, and facilitate estimation for disease outbreaks with under-reporting for which this framework can be applied. We assume that each individual, on becoming removed, is reported immediately with probability p independently among removals. We note here that in many practical situations reporting precedes removal, in which case the reporting of a case can also be regarded as the removal, since we assume that action such as isolation (or culling in animal populations) takes place immediately. Specification of the model for full inference thus requires five states, involving the following quantities: $I_r(t)$ – the number of infectious individuals at time t that will become reported cases ; $I_u(t)$ – the number of infectious individuals at time t that will not be reported; $R_r(t)$ and $R_u(t)$ – the number of reported and unreported removals respectively at time t; and $S(t)$ – the total number of susceptibles at time t. Hence we have the total number of infectious

and removals at time t which satisfy

$$I(t) = I_r(t) + I_u(t), \quad R(t) = R_r(t) + R_u(t) \qquad \text{for all } t \geq 0. \tag{3}$$

For the purpose of developing methods that use information resulting solely from reported cases, inference here is based on $S(t), I_r(t), R_r(t)$, with the transition probability in (1) replaced by

$$\Pr\{S(t+dt) = S(t) - 1,\ I_r(t+dt) = I_r(t) + 1\} = \beta S(t) p I(t)\, dt + o(\, dt). \tag{4}$$

Equation (2) can then be written as

$$\Pr\{I_r(t+dt) = I_r(t) - 1,\ R_r(t+dt) = R_r(t) + 1\} = \gamma I_r(t)\, dt + o(\, dt), \tag{5}$$

where γ now gives the rate of removal per reported case. Initial conditions under this model are given in Section 2.2.

We note that the distinct separation between the underlying epidemic process and the reporting/observation process is central to inference in the modelling framework developed here. Splitting both I and R classes in reported and unreported cases implies that reporting does not influence the disease dynamics, which facilitates identifiability between relevant parameters in the inference process.

When the 5-state model is explicitly considered, the full likelihood of the model can be written and RJMCMC employed to implement full statistical procedures as in Gamado *et al.* (2014). However, RJMCMC suffers efficiency issues once N reaches a few hundreds, and is not really feasible if N is in the thousands. As this becomes extremely time consuming (or potentially prohibitive) for increasing N, we investigate relevant approximations based on (4) and (5).

2.2 Approximation 1

We henceforth assume a large population size, N, and make use of the following approximations. For given reporting probability p, we assume, due to independence of individuals and constant rate of reporting, that

$$I_r(t) \sim \text{Bin}(I(t), p) \text{ and } R_r(t) \sim \text{Bin}(R(t), p). \tag{6}$$

Considering the mean of the binomial distribution, the assumptions in (6) imply, according to the law of large numbers, that the number of reported infectious and removals at time t can be approximated as the expected values

$$I_r(t) \approx pI(t) \text{ and } R_r(t) \approx pR(t). \tag{7}$$

In the first method of inference developed in this paper, we focus on the approximation that uses (7). Used independently of the binomial assumptions in (6), the approximations in Equations (7) correspond to the assumption that exactly a proportion p of the infective cases are reported. For large epidemics, the approximations (7) turn out to be accurate using asymptotic arguments. The more infections that exist, the closer the reported proportion tends to be p times the true number of infections. However, the validity of these approximations when the number of infections is small (e.g. at the beginning of the epidemic) is also important since we use them throughout the evolution of the epidemic. At the early stages of the epidemic, using (7) will likely underestimate the variability resulting from the reporting process, as shown in the next two sections. This motivates the adjustment on the reporting probability p in Section 2.3.

Hence using (7) in (4), we have the approximation

$$\Pr\{S(t+dt) = S(t) - 1,\ I_r(t+dt) = I_r(t) + 1\} \approx \beta S(t) I_r(t)\, dt + o(\, dt). \tag{8}$$

The infection rate β in the equation above can be interpreted as the contact rate between reported infected cases and susceptible individuals. Equation (8) implies that infections corresponding to reported cases occur approximately at rate $\beta S(t) I_r(t)$. The assumption of closed population implies that the population size is $N = S(t) + I(t) + R(t)$. Hence making use of Equation (7) we approximate $S(t)$ by

$$S(t) \approx N - \frac{I_r(t) + R_r(t)}{p}. \tag{9}$$

The two approximations (8) and (9) put together lead to the probability

$$\Pr\{S(t+dt) = S(t) - 1,\ I_r(t+dt) = I_r(t) + 1\} \approx \beta I_r(t) \left(N - \frac{I_r(t) + R_r(t)}{p} \right) dt + o(\, dt). \tag{10}$$

The initial conditions under our approach are $I_r(0) = 1$ and $S(0) \approx N - 1/p$.

2.3 Approximation 2

The approximations in (7) are quite restrictive regarding the variability of p, since they only use the mean of the binomial distributions in (6). Therefore (7) can be regarded as an approximation involving $\hat{p} = K_r/K$ (which is exactly the case when $T = \infty$) and thus Equation (15) is better regarded as a likelihood involving $\hat{p}$ rather than p. To allow for the variability about p, the actual proportion reported, to be taken into account in the likelihood, we use the assumption that the number of reported cases is binomial, i.e.

$K_r \sim \text{Bin}(K, p)$. The binomial assumption implies, based on the central limit theorem, that

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{K}\right),$$

and because K can be estimated by $K = K_r/p$, we obtain

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p^2(1-p)}{K_r}\right). \tag{11}$$

2.4 Approximation 3

2.4.1 Final size distribution

Given the contact rate β and the distribution of the infectious period, the distribution of the final size can be obtained by solving a system of equations that we introduce in this section. Let $\psi(\theta) = \boldsymbol{E}\left[\exp(-\theta\mathbb{I})\right]$ be the moment generating function of the infectious period $\mathbb{I}$, and P_N^k the probability that the final size of the epidemic is equal to k where $0 \leq k \leq N$. Then, P_N^k satisfies the following triangular system of equations (Andersson and Britton, 2000):

$$\sum_{k=0}^{l} \frac{\binom{N-k}{l-k} P_N^k}{\left[\psi\left(\beta(N-l)\right)\right]^{k+a}} = \binom{N}{l}, \quad \text{for } l = 0, \ldots, N, \tag{12}$$

where N is the initial number of susceptibles in the population and a is the initial number of infectives.

Numerical solutions of (12) can lead to negative probability values due to rounding errors. Demiris and O'Neill (2006) discuss that even for moderate population sizes greater than 100, these numerical problems occur with certainty. To avoid numerical problems, the approach proposed in Demiris and O'Neill (2006) is multiple precision arithmetic, which is computationally costly and time consuming. A quicker way to obtain an estimate of the final size is to use the following Gaussian approximation.

2.4.2 Gaussian approximation

Assume that we have a sequence of Generalised Stochastic Epidemics indexed by the initial susceptible population size N. Let K be the final size of the N^{th} epidemic and τ be the asymptotic proportion of individuals ultimately infected, i.e. $\tau = \lim_{N\to\infty} \frac{K}{N}$. Then τ is almost surely a constant as discussed by Andersson and Britton (2000), and in the case $R_0 > 1$, τ is the non-trivial solution of the non-linear equation

$$1 - \tau = \exp(-R_0\tau). \tag{13}$$

A general proof for this result is given in Andersson and Britton (2000). We can interpret this result by considering the left and right hand sides of the equation separately. The probability of escaping infection, for an individual faced by τ attacks, each affecting on average R_0, is equal to a 0 occurrence for the Poisson(τR_0) distribution, i.e. the right hand side of (13). On the other hand, the probability of escaping infection is equal to the proportion of initial susceptibles who remain uninfected, i.e the left hand side of (13).

If we let $\rho = 1 - \tau$, $\omega = \boldsymbol{E}(\mathbb{I})$ and $\varsigma^2 = \text{Var}(\mathbb{I})$, then for large N, the distribution of K is approximately Gaussian (Andersson and Britton, 2000):

$$K|R_0 \sim \mathcal{N}\left(\tau N, \frac{N(\rho(1-\rho) + (\beta N)^2 \varsigma^2 \tau \rho^2)}{(1 - \beta N \omega \rho)^2}\right), \tag{14}$$

where τ is the relevant solution of (13) obtained using Newton-Raphson method (Ortega and Rheinboldt, 2000). Demiris and O'Neill (2006) explore this approximation and compare it with the multiple precision arithmetic method and validate it for population sizes above 100.

3 Inference

3.1 Method 1

The new dynamic processes using the approximations (7), (8)-(10) lead to the approximate likelihood

$$\begin{aligned} L(\beta, \gamma, p; \boldsymbol{s_r}, \boldsymbol{r_r}) \approx & \prod_{i \in \mathcal{I}_{-w}} \beta I_r(s_i^-) \left(N - \frac{I_r(s_i^-) + R_r(s_i^-)}{p}\right) \\ & \exp\left(-\beta \int_{s_w}^{T} I_r(t) \left(N - \frac{I_r(t) + R_r(t)}{p}\right) dt\right) \\ & \prod_{i \in \mathcal{R}_r} \gamma \exp\left(-\gamma(r_i - s_i)\right) \end{aligned} \tag{15}$$

where $\boldsymbol{s_r} = (s_1, \ldots, s_{K_r})$ and $\boldsymbol{r_r} = (r_1, \ldots, r_{K_r})$ are respectively the vector of infection and removal times for reported cases, with K_r being the total number of reported cases ($K_r = R_r(T)$, T being the end of the observation period); w is the individual turning out to be the first infected among the reported removals, $\mathcal{I}_{-w}$ denotes the set of all the reported infective individuals excluding w and $\mathcal{R}_r$ is the set of reported removals.

It is expected when using likelihood (15) that good knowledge of p will lead to good estimation of β. But in practice, p is also of interest. The approximate likelihood function (15) will be used to estimate β, γ and p and inference made using (15) will be referred to as Method 1.

The likelihood in (15) is derived using the rate at which infections from reported cases occur and is a function of β, γ and p. Therefore direct estimation of the three model parameters can be made in a Bayesian framework allowing for incorporation of prior information in the analysis.

If we assume the following conjugate prior distributions for β and γ

$$\beta \sim \text{Ga}\left(\alpha_\beta, \nu_\beta\right) \text{ and } \gamma \sim \text{Ga}\left(\alpha_\gamma, \nu_\gamma\right), \tag{16}$$

where α is the shape and ν is the rate of the gamma distribution, we obtain the conditional posterior distributions

$$\beta|\gamma, p, \boldsymbol{s_r}, \boldsymbol{r_r} \sim \text{Ga}\left(K_r + \alpha_\beta - 1, \int_{s_w}^{T} I_r(t)\left(N - \frac{I_r(t) + R_r(t)}{p}\right)dt + \nu_\beta\right) \tag{17}$$

and

$$\gamma|\beta, p, \boldsymbol{s_r}, \boldsymbol{r_r} \sim \text{Ga}\left(K_r + \alpha_\gamma, \sum_{i=1}^{K_r}(r_i - s_i) + \nu_\gamma\right). \tag{18}$$

We assume a beta prior $\text{Beta}(\alpha_p, \tau_p)$ for p but cannot obtain a beta posterior distribution. This is mainly due to the fact that p is involved in the integration part of the likelihood (15). However, we can still sample from the posterior distribution of p using a Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970). It then remains to update the infection times of reported cases. We select one individual at random (say k), and propose an infection time, s_k, using the model assumption $r_k - s_k \sim \text{Exp}(\gamma)$. The proposed infection time is accepted with probability

$$A = \min\left(\frac{L^{\text{new}}}{L^{\text{old}}} \times \frac{\exp\{-\gamma(r_k - s_k')\}}{\exp\{-\gamma(r_k - s_k)\}}, 1\right), \tag{19}$$

where s_k' is the current infection time of individual k and L^{old} and L^{new} are the likelihood (15) evaluated at the appropriate parameter values. Pseudo-code detailing the algorithm for Method 1 is given in Appendix A.2.1.

3.2 Method 2

Our underlying Bayesian model can now be fully specified using the approximate likelihood function $L(\beta, \gamma, \hat{p};\, \boldsymbol{s_r}, \boldsymbol{r_r}) \times \phi(\hat{p} - p)$, where $L(\cdot)$ is given in Equation (15) with p replaced by $\hat{p}$ and $\phi(\cdot)$ is the density of the normal distribution with mean 0 and variance $p^2(1-p)/K_r$, together with the priors for β, γ and p from Section 3.1. Treating $\hat{p}$ (essentially K, since K_r is observed) as an auxiliary variable and given $\boldsymbol{s_r}$ and $\boldsymbol{r_r}$, the joint posterior density of β, γ, p and $\hat{p}$ is approximately given as

$$\pi_\beta(\beta)\pi_\gamma(\gamma)\pi_p(p)\phi(\hat{p} - p)L(\beta, \gamma, \hat{p};\, \boldsymbol{s_r}, \boldsymbol{r_r}), \tag{20}$$

where π_β, π_γ and π_p are the three prior densities on β, γ and p respectively. Equation (20) can be used to show the link between the posterior distributions of p and $\hat{p}$ (see Appendix A.1). A detailed algorithm for Method 2 in the form of pseudo-code is available in Appendix A.2.2.

3.3 Method 3

An alternative estimation approach is to employ an approximate Gibbs sampling scheme to obtain inference about the parameters β, γ and p. We refer to this as Method 3. Sampling from the posterior distributions of β and γ given respectively (γ, p) and (β, p) is straightforward using the approximate densities (17) and (18). To sample from the posterior distribution of p given β and γ, we consider the final size of the epidemic as an auxiliary variable and use results from the literature to estimate its distribution. We describe this method in the following sections.

3.3.1 Estimation of p given β and γ

In what follows, estimation of p in the case of a completed epidemic will require estimation of the final size $K = R(\infty)$. With the assumptions (7) and (9) leading to (15), we further assume that all available information about p is contained in the reported final size. Here, knowledge of β and γ is equivalent to knowledge of the reproduction number $R_0 = (\beta/\gamma)(N-1)$. Given the reproduction number, the final size of the epidemic can be estimated as we will present below. Once the final size is estimated given β, we can obtain a binomial estimate of p using

$$K_r \sim \text{Bin}(K, p). \tag{21}$$

The term in the binomial approximation (21) that contains p is of the form

$$p^{K_r}(1-p)^{K-K_r}, \tag{22}$$

and hence, assuming a $\text{Beta}(\alpha_p, \tau_p)$ prior on p we can sample from its conditional posterior distribution

$$p|K, \boldsymbol{s_r}, \boldsymbol{r_r}, \beta \sim \text{Beta}(K_r + \alpha_p, K - K_r + \tau_p). \tag{23}$$

Given β and γ, we can then sample in turn K from the Gaussian approximation (14) and p from the beta distribution (23).

3.3.2 Approximate Gibbs sampling algorithm

Our aim is to run a Gibbs sampler for the joint posterior distribution of β, γ and p. We have already determined the full conditional posterior distributions of β and γ in (17) and (18). To update p conditional on (β, γ), i.e. R_0, we sample from the conditional distribution of K, the prior distribution on p and the data, where actually the prior on p and the data are relatively non-informative since the conditional density of K in (14) does not involve p and the reported times. Therefore, we can just sample from the conditional distribution of K given R_0 under the constraint $K \geq K_r$. Given K (and β which is not directly involved here), we can sample p using the full conditional posterior distribution (23). A full description of the algorithm for Method 3 in the form of pseudo-code is presented in Appendix A.2.3.

4 Application to simulated data

We apply the methods developed in Section 3 to simulated data. As the interest of our methods mainly lies in their applicability to the estimation of large-scale and fast-developing epidemics, we consider cases where R_0 is greater than 1. With recent Ebola cases providing examples of such epidemics, our application data are generated with R_0 being consistent with estimates from previous and recent Ebola outbreaks.

More specifically, we first consider a population of size $N = 600$ and simulate an epidemic based on the two processes described with Equations (1) and (2). We choose R_0 using relevant literature estimates in Ebola cases. For instance, estimates of R_0 from outbreaks in Congo (1995) and Uganda (2000) ranged from 1.3 (Chowell *et al.*, 2004) to 2.7 (Legrand *et al.*, 2007). Maximum likelihood estimates of the basic reproduction number in the 2014 Ebola Virus disease in Guinea, Sierra Leone and Liberia ranged between 1.51 and 2.53 (Althaus, 2014). Therefore, we consider $R_0 = 2$ and, aiming at flexibility and more general applicability, we set $\gamma = 1$ and scale the contact parameter (according to R_0 and N) to be $\beta = 3.333 \times 10^{-3}$. These values imply a number of $3.333 \times 10^{-3} \times 600 = 2.0$ effective contacts per time unit and average length of infectious period equal to 1 time unit. In addition, Merler *et al.* (2015) consider an under-reporting scenario with reporting rate 50% in the general population, and their analysis leads to an estimate of $R_0 = 1.9$ (95% CI $1.62 - 2.14$). We therefore follow the same reporting assumption, and consider the true reporting probability to be $p = 0.5$.

With this parameter setting, a realisation of the simulated epidemic gave $K_r = 200$

reported infections out of a typical large outbreak of $K = 414$ ultimately infected cases, as shown in Table 1 together with the model parameters.

parameters					outcome	
N	β	γ	R_0	p	K	K_r
600	0.0033	1.0	2.0	0.5	414	200

Table 1: True parameters values and outcome of a single realisation of the model simulation

4.1 Inference using Methods 1-3

We estimate all model parameters under imperfect reporting implementing Metropolis-Hastings-within-Gibbs algorithms as described in Sections 3.1, 3.2 and 3.3 for Methods 1, 2 and 3 respectively.

The priors on β and γ are chosen as non-informative $\text{Ga}\,(0.001, 0.001)$ distributions, while for p we assume a non-informative $\text{Beta}(1, 1)$ distribution. The posterior estimates are summarised in Table 2 for all the parameters under the different methods.

Estimates of the posterior mean of the parameters are quite close to the true values in all three methods, particularly for β, γ and p. For the epidemic size parameter, K, Method 2 seems to be more accurate, with Method 3 performing less well. Credible intervals for all four parameters under the three methods comfortably contain the true parameter values. However, Method 3 gives higher variance estimates for β, p and K, resulting in higher uncertainty as also demonstrated by wider credible intervals.

These results are also illustrated in Figures 1(a)-1(d), where we superimpose the marginal posterior densities of the parameters with Methods 1, 2 and 3. The vertical green solid line represents the true value for each parameter. The true values for β and γ fall well within their marginal posterior densities under all methods. The distributions of K and p are respectively left-skewed and right-skewed, with the right-skewness of p being directly related to the left-skewness of K, since small values of K lead to large values of p. However the true values are close to the mode of their densities. Overall the plots demonstrate that the approximation techniques seem to work well and we are able to recover the true parameter values.

The convergence and mixing of the Markov chains were assessed by inspecting the chain traces, the auto-correlation functions and correlation between parameters. There was no evidence of lack of convergence. The auto-correlation functions of β and p are plotted on

Figures 2 and 3 respectively, under Methods 1-3. The auto-correlation functions of γ and K are provided in Appendix A.4. Method 3 appears to provide lower auto-correlation in the MCMC chain, demonstrating better mixing of the algorithm.

4.2 Comparison with a full RJMCMC scheme

The methods in Section 3 involve approximations, while a method based on the exact likelihood and RJMCMC has been previously proposed in Gamado *et al.* (2014). The RJMCMC method involves imputing the missing infection and removal times of the unreported individuals along with the infection times of the reported cases, as detailed in the algorithm in Appendix A.3. However, the RJMCMC approach is highly time consuming for large populations. We therefore present here a comparison of the results using RJMCMC and the approximate methods developed in this paper.

We apply Methods 1 - 3 and the RJMCMC approach to the simulated epidemic data described in Section 4 (population size $N = 600$, parameters given in Table 1), running each method for $n = 100000$ MCMC iterations with 10000 burn-in periods. The posterior distributions obtained under the RJMCMC scheme (referred to as Method 4) are added to those under Methods 1- 3, as summarised in Table 2. A graphical demonstration of the new results is also given in Figures 1(a), 1(b), 1(c) and 1(d) for β, γ, p and K respectively.

The estimates in Table 2 and Figures 1(a) - 1(d) demonstrate that the approximate methods developed here produce similar results to those obtained when using RJMCMC. Figures 1(a) - 1(d) suggest that while Method 3 provides good estimates of the posterior means, estimates for p and K are not as close to the RJMCMC results as those from Methods 1 and 2. Method 3 also seems to be overestimating the posterior variances of the model parameters except γ. This may be explained by the fact that γ is not directly involved in the approximations.

We also provide in Table 2 the effective sample size of the chains under the four methods, defined in Kass *et al.* (1998) as

$$ESS = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

where n is the total simulated chain size and ρ_k is the autocorrelation function at lag k. The computations were performed using CODA (Plummer *et al.*, 2006). It is clear that Method 3 provides the highest effective sample size for three out of the four parameters, giving an indication of better mixing chains. This is also confirmed by the average effective sample size over all four parameters, shown in Table 3. The table shows that Method 1 is the next best performer in terms of ESS, followed by Method 2.

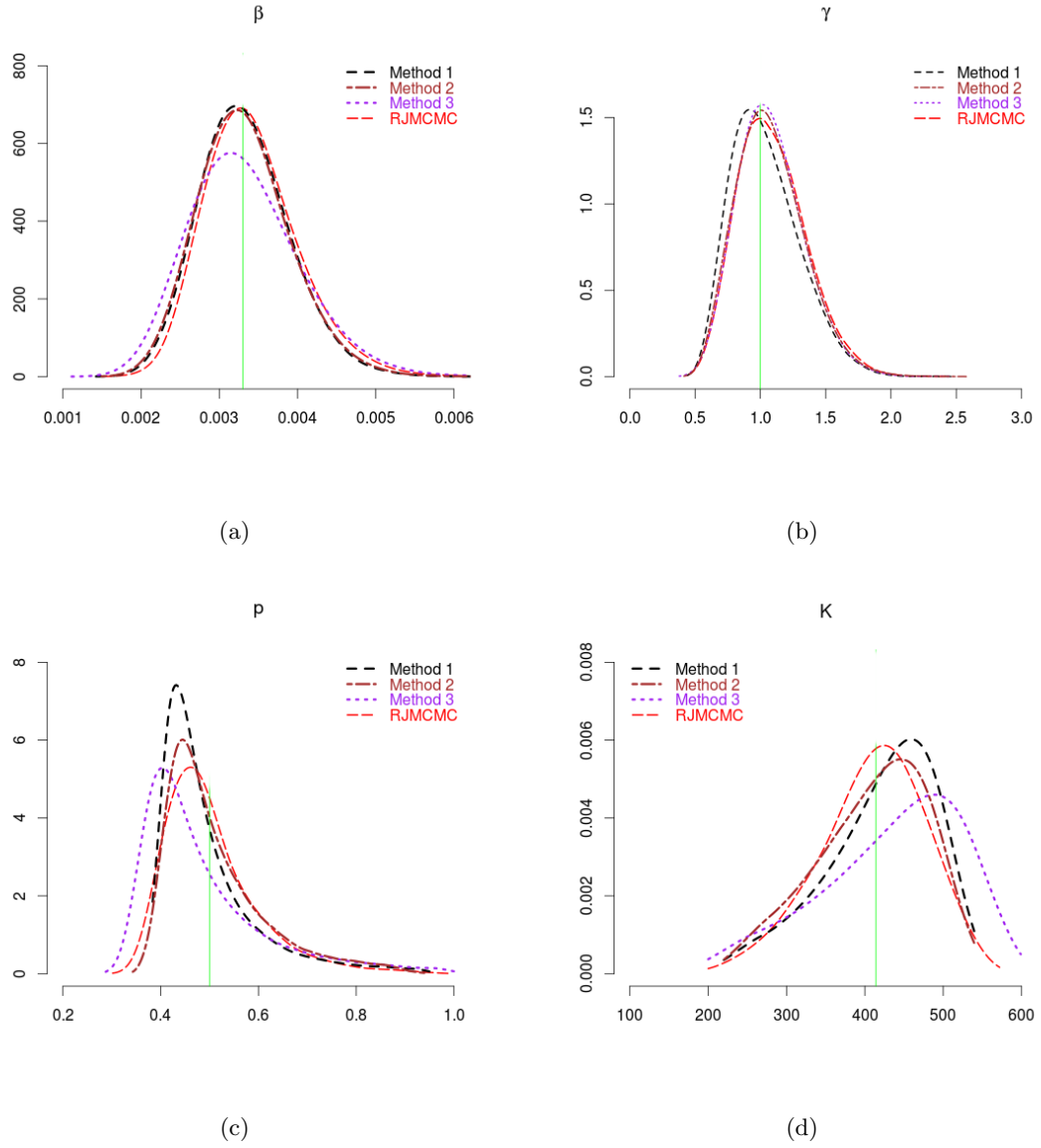
		mean	sd	C.I.	ESS
β	Method 1	0.00333	0.00057	$(0.00232, 0.00454)$	1487.925
	Method 2	0.00332	0.00059	$(0.00229, 0.00459)$	973.594
	Method 3	0.00332	0.00070	$(0.00214, 0.00487)$	1245.583
	Method 4	0.00341	0.00060	$(0.00243, 0.00476)$	518.291
γ	Method 1	1.038	0.259	$(0.628, 1.612)$	387.480
	Method 2	1.077	0.259	$(0.645, 1.644)$	331.812
	Method 3	1.079	0.252	$(0.649, 1.626)$	416.875
	Method 4	1.090	0.261	$(0.652, 1.662)$	288.976
p	Method 1	0.496	0.099	$(0.397, 0.795)$	719.406
	Method 2	0.511	0.103	$(0.390, 0.798)$	947.489
	Method 3	0.485	0.127	$(0.344, 0.849)$	1387.515
	Method 4	0.502	0.102	$(0.373, 0.787)$	447.618
K	Method 1	420.184	69.111	$(255.628, 518.436)$	580.293
	Method 2	408.153	69.788	$(254.423, 514.700)$	638.307
	Method 3	434.722	88.117	$(235.189, 561.964)$	1172.351
	Method 4	409.812	65.369	$(265.000, 521.000)$	251.904

Table 2: Posterior estimates of β, γ, p and K when applying Methods 1, 2, 3 and 4 (RJMCMC) to the same dataset: 200 reported cases with $\beta = 0.0033, \gamma = 1$ ($R_0 \approx 2$) and $p = 0.5$

	Method 1	Method 2	Method 3	Method 4
Average ESS	793.8	722.8	1055.6	376.7
CPU time	$47,361$	$49,601$	$3,438$	$361,727$

Table 3: Average ESS and CPU times (in seconds) for running the four methods for the same data

Table 3 also presents the record of the CPU times (in seconds) for running each of the methods on the same computer. Method 3 takes significantly less time to complete, outperforming the slowest method (Method 4, RJMCMC) by a factor of about 100. The computing times for Methods 1 and 2 are very similar with a slight advantage to Method 1. It is worth noticing that all three approximate methods are considerably more efficient than the RJMCMC approach (Method 4) in terms of both ESS and CPU time, with



(a) (b)

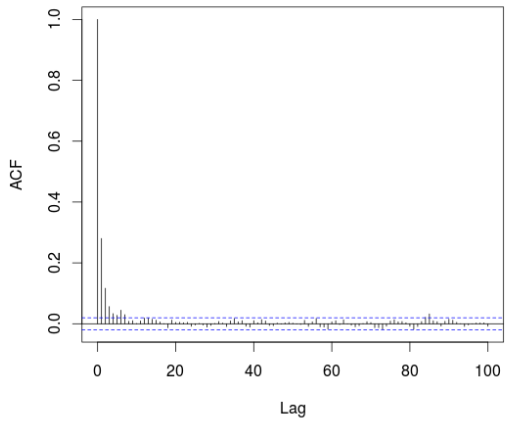
(c) (d)

Figure 1: Posterior densities of the parameters β, γ, p and K with $N = 600$ data ((a), (b), (c) and (d) respectively) when using Method 1 (black dashed line), Method 2 (brown two-dashed line), Method 3 (purple dotted line) and RJMCMC (red long-dashed line)

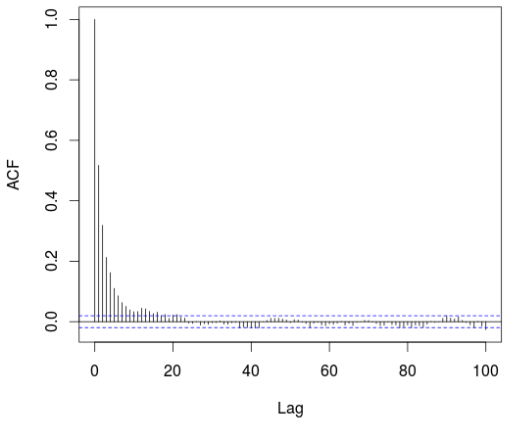
Method 3 being the most efficient overall.

Considering the auto-correlation of the chains for β and p in Figures 2 and 3 (and similar plots for γ and K in Appendix A.4), Method 3 again provides better mixing, followed by Method 1. The RJMCMC method has the highest auto-correlation, therefore suggesting slower mixing.

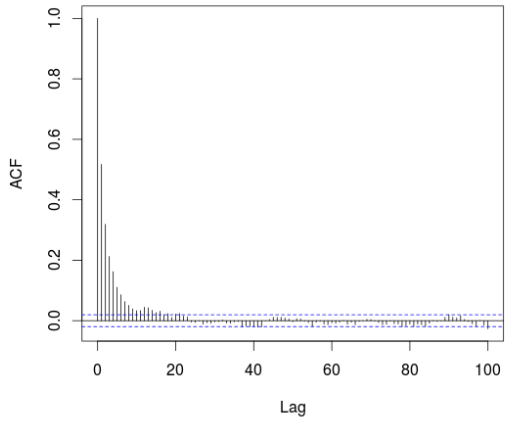
The results in this section clearly suggest that the approximation-based methods developed in this paper are similar in accuracy to the RJMCMC approach, while also being more efficient in terms of chain mixing and real time running requirements.



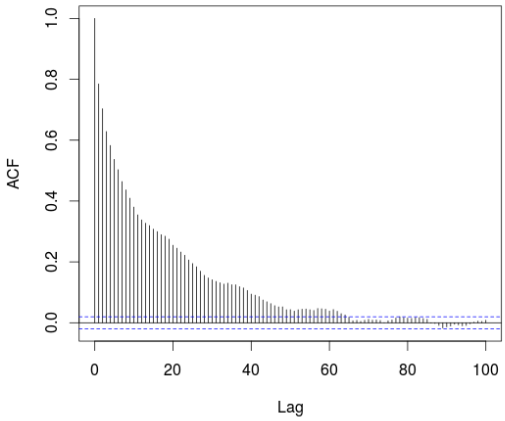
a) ACF plot for β, Method 1



b) ACF plot for β, Method 2



c) ACF plot for β, Method 3



d) ACF plot for β, Method 4

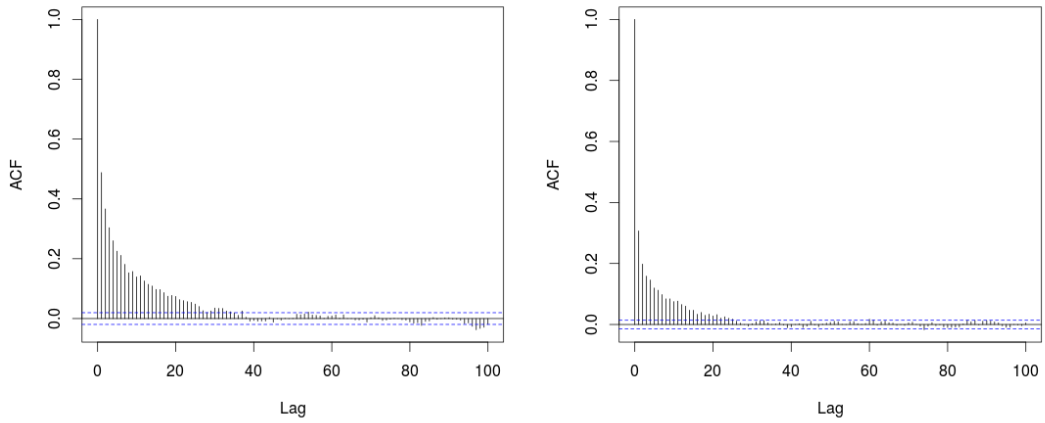
Figure 2: Auto-correlation function for parameter β (data from Section 4)

5 Simulation Studies

5.1 Population size $N = 600$

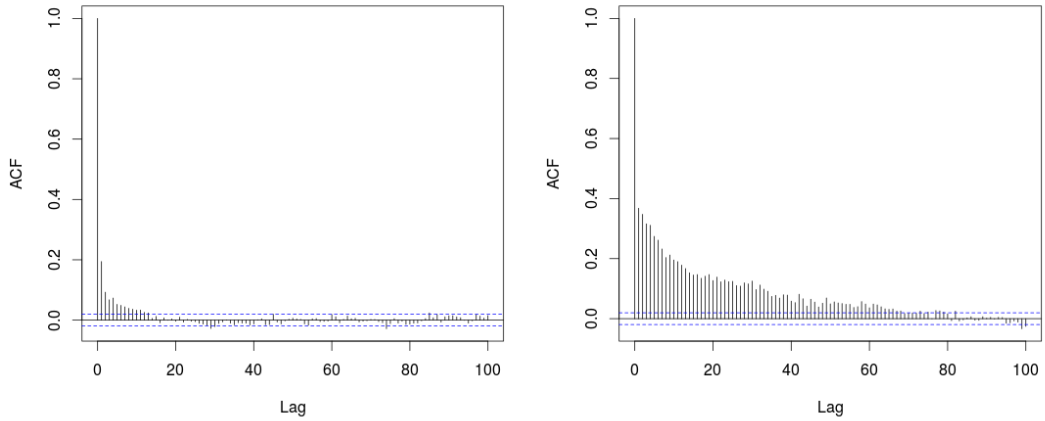
We now perform an extended simulation study for all Methods 1, 2, 3 and 4. We simulate $N_s = 1000$ epidemic outbreaks using the parameter values specified in Table 1, and for each simulated data set we make inference for the parameters β, γ, p, R_0 and K, after $n = 5000$ MCMC iterations with 1000 burn-in periods. The posterior mean, standard deviation, median and credible intervals are monitored for each parameter. Average results from the 1000 simulations are shown in Table 4, where the standard error of the mean (i.e. the standard deviation of the estimator) is also given in brackets.

We also present in Table 4 the coverage rate of each 95% credible interval, and the relative mean squared error (RMSE) of the estimators of the parameters under the four



a) ACF plot for p, Method 1

b) ACF plot for p, Method 2



c) ACF plot for p, Method 3

d) ACF plot for p, Method 4

Figure 3: Auto-correlation function for parameter p (data from Section 4)

methods. The RMSE is a scale-free measure defined for a parameter θ with estimator $\hat{\theta}$ as

$$\text{RMSE}_\theta = \boldsymbol{E}\left(\frac{\hat{\theta}-\theta}{\theta}\right)^2. \tag{24}$$

The results in Table 4 confirm that the three approximate methods developed here perform well in estimating the model parameters, with all mean estimates being very close to the true values. As with the single simulated data set, Method 3 appears to overestimate the variance, compared to the other methods. We note that the coverage rate of the credible intervals for K, p and R_0 under Method 3 is particularly high, again pointing to the suggestion that this method overestimates the uncertainty associated with these parameter estimates. Method 2 performs generally better than Method 1, with slightly higher coverage rate for p and K and lower RMSE.

The RMSE for all parameters is small demonstrating once again that all three approxi-

		mean (s.e.)	sd	CI cov. rate	RMSE
β	Method 1	3.25×10^{-3} (1.45×10^{-5})	4.84×10^{-4}	93.9%	0.022
	Method 2	3.24×10^{-3} (1.50×10^{-5})	4.87×10^{-4}	93.3%	0.024
	Method 3	3.160×10^{-3} (1.50×10^{-5})	6.62×10^{-4}	97.8%	0.021
	Method 4	3.28×10^{-3} (1.48×10^{-5})	4.86×10^{-4}	93.7%	0.021
γ	Method 1	9.45×10^{-1} (6.67×10^{-3})	2.05×10^{-1}	87.1%	0.050
	Method 2	9.54×10^{-1} (7.43×10^{-3})	2.05×10^{-1}	87.1%	0.046
	Method 3	9.80×10^{-1} (6.38×10^{-3})	2.36×10^{-1}	95.8%	0.056
	Method 4	1.08×10^{0} (7.45×10^{-3})	2.26×10^{-1}	91.0%	0.048
p	Method 1	4.95×10^{-1} (2.08×10^{-3})	7.32×10^{-2}	89.2%	0.039
	Method 2	4.99×10^{-1} (2.05×10^{-3})	7.75×10^{-2}	91.9%	0.035
	Method 3	5.37×10^{-1} (1.73×10^{-3})	1.12×10^{-1}	99.6%	0.019
	Method 4	5.11×10^{-1} (2.05×10^{-3})	7.73×10^{-2}	93.4%	0.017
K	Method 1	482.676 (1.69)	58.891	91.3%	0.026
	Method 2	480.327 (1.70)	57.540	91.9%	0.014
	Method 3	455.860 (1.29)	75.056	99.6%	0.033
	Method 4	469.754 (1.69)	57.483	91.8%	0.013
R_0	Method 1	2.145 (1.88×10^{-2})	0.559	94.0%	0.099
	Method 2	2.117 (1.89×10^{-2})	0.545	94.0%	0.097
	Method 3	2.016 (6.23×10^{-3})	0.421	99.9%	0.101
	Method 4	2.058 (1.88×10^{-2})	0.544	93.3%	0.093

Table 4: Simulation study estimates using: (a) Method 1; (b) Method 2; (c) Method 3; and (d) Method 4. The average number of reported cases over the 1000 simulations was $K_r = 236.35$, while the true average final size was $K = 472.25$

mate methods are performing well when both the bias and the variability of the estimator (standard error) are taken into account. Method 4 has the smallest RMSE for all parameters. Comparing the approximate methods, Method 3 performs better for β and p, although the RMSE for the estimator of parameter β is very close using all methods. For p, it appears that the high bias under Method 3 is compensated by a lower standard error of the estimator, leading to low RMSE. On the other hand, Method 3 performs less well in estimating the final epidemic size of the epidemic, where the true average value over the 1000 simulations is $K = 472.25$.

The higher variability of the estimated final size under Method 3 may be explained by the fact that this approach uses an approximate distribution for the final epidemic size, K, in addition to the approximation the likelihood shares with the other two methods.

Table 5 contains relative CPU times for running the four methods for the simulation studies. Method 3 is the quickest to run and therefore, considering ratios with respect to the CPU time of Method 3, the table shows that Methods 1, 2 and 4 are respectively 28, 31 and 151 times slower than Method 3.

	Method 1	Method 2	Method 3	Method 4
Ratio of CPU time	28.36	31.20	1	150.76

Table 5: Relative comparisons (ratio) of CPU times for running Methods $1-4$ for the simulation studies with population size $N = 600$ and $R_0 = 2$

5.2 Basic reproduction number $R_0 = 1.5$

We also consider the performance of Methods 2 and 3 with a smaller value of R_0, but ensuring that adequate epidemic sizes are observed. Method 3 provides biased estimates for values of R_0 close to 1, due to the constraint $R_0 > 1$ (necessary to obtain a non-zero solution for Equation (13)), implying that any set of proposed parameter values leading to $R_0 < 1$ is automatically rejected in the MCMC process. We have therefore considered the case $R_0 = 1.5$, which gives a small such rejection rate (4.86%). For consistency with earlier simulations, $R_0 = 1.5$ suggests a population size of $N = 810$ for obtaining an average epidemic size of $K = 472.18$ with $K_r = 236.98$ reported cases. The results are shown in Table 6, and are consistent with the estimates and findings in the case of $R_0 = 2$. Both approximations estimate well the model parameters, with Method 3 giving larger variance estimates with higher coverage rates for all parameters, and lower RMSE for p.

5.3 Simulation studies with varying population size

The approximations in the methods developed here rely on a large population with an epidemic taking off. A study of how well they work with respect to the size of the population size, is therefore also of interest. We focus on Methods 2 and 3, since the analysis in previous sections suggests that Methods 1 and 2 exhibit very similar performance, with Method 2 showing a slight advantage. We run simulations with different population sizes by choosing the contact rate β such that the reproduction number R_0 is unchanged ($R_0 \approx 2$). For different population sizes $N \in \{300, 600, 900, 1200, 1500, 2000, 4000\}$, the contact rate β

		mean (s.e.)	sd	CI cov. rate	RMSE
	Method 2	1.84×10^{-3} (9.06×10^{-6})	2.95×10^{-4}	92.1%	0.018
β	Method 3	1.86×10^{-3} (9.08×10^{-6})	3.38×10^{-4}	97.2%	0.017
	Method 2	9.75×10^{-1} (5.99×10^{-3})	1.92×10^{-1}	91.0%	0.045
γ	Method 3	9.51×10^{-1} (6.07×10^{-3})	2.03×10^{-1}	97.5%	0.055
	Method 2	5.03×10^{-1} (2.18×10^{-3})	1.03×10^{-1}	90.3%	0.032
p	Method 3	4.83×10^{-1} (1.75×10^{-3})	1.34×10^{-1}	100%	0.018
	Method 2	463.16 (2.21)	84.93	90.0%	0.021
K	Method 3	482.40 (1.82)	109.01	100%	0.029
	Method 2	1.50 (1.56×10^{-2})	0.29	92.9%	0.099
R_0	Method 3	1.53 (1.21×10^{-2})	0.34	99.8%	0.108

Table 6: Simulation study estimates using Method 2 and Method 3 in the case of true parameter values $\beta = 1.85 \times 10^{-3}$, $\gamma = 1$ and $R_0 = 1.5$ on a population size of $N = 810$. The average number of reported cases over the 1000 simulations was $K_r = 236.98$, while the true average final size was $K = 472.18$

was chosen to be respectively $\beta \in \{0.0067, 0.0033, 0.0022, 0.00167, 0.0013, 0.001, 0.0005\}$. For each of the simulation cases, we consider the mean of the posterior distribution as a point estimate and estimate the RMSE of the estimators of parameters β, γ and p, as defined in Equation (24).

We plot the RMSE as a function of the population size in Figure 4. For both methods there is a general decrease in the RMSE of all parameter estimators as the population size increases. The plot does not give evidence against asymptotically zero error for the two methods: the larger the population size is, the more accurate the estimations become. The plot also shows that Method 3 is more accurate for parameters β and p in smaller populations, but this advantage wears off as the size of the population becomes larger and asymptotic approximations work better. This trend seems to reverse for parameter γ, which may not be surprising as its estimation relies less on the employed approximations.

6 Discussion

In this paper we have considered the SIR epidemic model with constant probability of reporting. Departing from earlier approaches, which can be implemented through RJM-

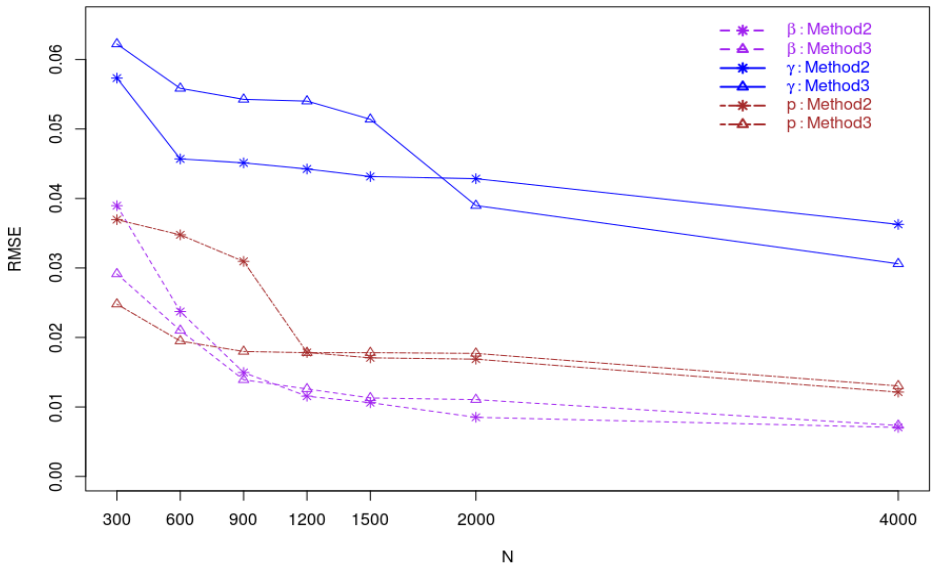


Figure 4: Comparison of RMSE of estimators of β, γ and p under Methods 2 and 3 for varying population size N. Parameter β: dotted (violet) line; γ: solid (blue) line; p: dashed (brown) line. Method 2: lines with superimposed asterisks; Method 3: superimposed triangles

CMC, we have derived an approximate likelihood, based on which we have been able to propose three different methods to estimate the model parameters. The first method involves use of approximation by assuming a fixed fraction of cases being reported, while the second method allows for randomness in which removals are reported but approximating the underlying binomial distribution by a normal. A third method has utilised an approximate Gibbs sampling algorithm for fast inference on the problem. All three methods give a good solution to the parameter estimation problem, with the Gibbs sampling approach (Method 3) being more efficient in terms of actual running time and ESS, at the cost of overestimating the variance of the parameters to a small degree.

The quality of the estimation was demonstrated with application on a generated data set, and was also confirmed with a simulation study where 1000 epidemic outbreaks were generated. All simulation scenarios were produced under realistic assumptions regarding the rate of spread of the disease, drawing on experience from recent Ebola epidemics. Inference based on the approximation methods was also compared to estimates obtained with a RJMCMC method (Gamado *et al.*, 2014) and turned out to closely agree.

In small-scale outbreaks the incomplete nature of available information can lead to confounding issues for parameters that drive the dynamics of the epidemic, and therefore estimation may suffer from identifiability problems between the contact rate β and the

reporting probability p. However, application of our methodology to moderate-to-large scale epidemics suggests that there may be sufficient information in the removal process to obtain accurate inferences, even in the presence of under-reporting.

Among the epidemiological factors that influence the reporting process is the severity of the exhibited symptoms, leading to asymptomatic cases being regarded as non-reported. In this paper we have assumed that infectiousness is the same for symptomatic and asymptomatic individuals and have treated the latter as typical unreported cases as opposed to work by Chowell *et al.* (2007) where a different compartment is considered. For the infectious periods, we have assumed here an exponential distribution to describe the uncertainty associated with their duration. However extensions to non-Markovian models (e.g. O'Neill and Becker, 2001; Streftaris and Gibson, 2004b, 2012) should be possible.

A principal advantage of the approximate methods presented in this paper is the time cost and mixing efficiency of the associated algorithms. Run on the same computer, all three approximate methods performed faster than the RJMCMC based approach. RJMCMC methodology is recommended when the population size is not too large, in which case more accurate results can be obtained. In general, as expected, gains in efficiency tend to occur at the expense of accuracy. Nevertheless, the most time-efficient option amongst those considered here, would be to use Method 3, involving the Gibbs sampling approach, which overwhelmingly outperforms the RJMCMC method. Method 2 offers a slight improvement to accuracy, whilst still being considerably faster than the RJMCMC approach. Our approximation-based methodology shares conceptual similarities with particle learning algorithms, as described by Johannes *et al.* (2009), Lopes *et al.* (2011) and Dukic *et al.* (2009). Although the latter are applied to state-space models, not allowing direct comparisons, both approaches have the potential of being computational faster than other standard methodology.

The approximate methods developed in this paper provide useful tools in cases where the volume of data and urgency imposed by the rapid progress of an outbreak necessitate fast and reliable inference on the spread dynamics of the disease. Although the methods used here are applicable to completed epidemics, they can help inform decision making in situations of ongoing epidemics where historical data of similar or re-emerging pathogens may be used, or when local-scale outbreaks can be adequately modelled as completed within reasonable time windows, in order to infer epidemic dynamics on a wider scale. This can be central to the timely development of control policies and can also provide the basis of subsequent full characterisation of the epidemic determinants.

References

C.L. Althaus. Estimating the reproduction number of ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLOS Currents Outbreaks*, 2014. doi: 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.

H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer, 2000.

O. N. Bjornstad, B. F. Finkenstadt, and B. T. Grenfell. Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184, 2002.

I. Chis-Ster and N. M. Ferguson. Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS ONE*, 2(6):e502, 2007.

G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J Theor Biol*, 229(1):119–126, 2004.

G. Chowell, H. Nishiura, and L. M. Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J. R. Soc. Interface*, 4(12):155–166, 2007.

J. A. Clarkson and P. E. M. Fine. The efficiency of measles and pertussis notification in England and Wales. *International Journal of Epidemiology*, 14:153–168, 1985.

N. Demiris and P. D. O'Neill. Computation of final outcome probabilities for the generalised stochastic epidemic. *Statistics and Computing*, 16(3):309–317, 2006.

I. Dorigatti, S. Cauchemez, A. Pugliese, and N. M. Ferguson. A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: Application to the Italian 20092010 A/H1N1 influenza pandemic. *Epidemics*, 4(1):9–21, 2012.

V. M. Dukic, H. F. Lopes, and N. Polson. Tracking flu epidemics using google flu trends and particle learning. *SSRN*, 2009. `http://ssrn.com/abstract=1513705` or `http://dx.doi.org/10.2139/ssrn.1513705`.

C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut,

O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. Ma. Espejo Guevara, F. Checchi, E. Garcia, S. Hugonnet, and C. Roth. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324:1557–1561, 2009.

K. M. Gamado, G. Streftaris, and S. Zachary. Modelling under-reporting in epidemics. *Journal of Mathematical Biology*, 69(3):737–765, 2014.

G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics Applied in Medicine and Biology*, 15:19–40, 1998.

W. K. Hastings. Monte Carlo sampling using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

J. A. P. Heesterbeek and K. Dietz. The concept of R_0 in epidemic theory. *Statist. Neerlandica*, 50(1):89–110, 1996.

N. Hens, M. Van Ranst, M. Aerts, E. Robesyn, P. Van Damme, and P. Beutels. Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: a multi-country analysis for influenza A/H1N1v 2009. *Vaccine*, 29:896–904, 2011.

C.P. Jewell, M.J. Keeling, and G.O. Roberts. Predicting undetected infections during the 2007 foot and mouth disease outbreak. *JRS Interface*, 6:1145–1151, 2008.

M. Johannes, N. G. POLSON, and S. M. YAE. Particle learning in nonlinear models using slice variables. *Biometrika*, pages 1–17, 2009.

R. E. Kass, B. P. Carlin, A. Gelman, and R. Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, 52:93–100, 1998.

E.S. Knock and P. D. O'Neill. Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics (Oxford, England)*, 15(1):46–59, 2014.

J. Legrand, R.F. Grais, P.Y. Boelle, A.J. Valleron, and A. Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiol Infect*, 135(4):610–621, 2007.

H. F. Lopes, C. M. Carvalho, M. S. Johannes, and N. G. Polson. Particle learning for sequential bayesian computation. *Bayesian Statistics*, 9:317–360, 2011.

S. Merler, M. Ajelli, L. Fumanelli, M. F. C. Gomes, A. P. Y Piontti, L. Rossi, D. L. Chao, I. M. Longini Jr, M. E. Halloran, and A. Vespignani. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15 (2):204–211, 2015.

N. Metropolis, A. W Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

P. Neal and C. L. Huang. Forward simulation MCMC with applications to stochastic epidemic models. *Scandinavian Journal of Statistics*, 42(2):378–396, 2015.

P. Neal and G. Roberts. A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15:315–327, 2005.

P.D. O'Neill and N.G. Becker. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2, 1:99–108, 2001.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables (Classics in Applied Mathematics, 30)*. Philadelphia, PA: SIAM, 2000.

O. Papaspiliopoulos, G. O. Roberts, and M. Skold. Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics, 7 (Tenerife 2002)*, pages 307–326, 2003.

M. Plummer, N. Best, K. Cowles, and K. Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

G. Streftaris and G. Gibson. Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, 4:63–75, 2004a.

G. Streftaris and G.J. Gibson. Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proc. R. Soc. Lond. B*, 271:1111–1117, 2004b.

G. Streftaris and G.J. Gibson. Non-exponential tolerance to infection in epidemic systems – modelling, inference and assessment. *Biostatistics*, 13(4):580–593, 2012.

L. F. White and M. Pagano. Reporting errors in infectious disease outbreaks, with an application to pandemic influenza A/H1N1. *Epidemiologic Perspectives & Innovations*, 7, 2010.

WHO: World Health Organization. Guinea: The ebola virus shows its tenacity. 2015. Available from http://www.who.int/csr/disease/ebola/one-year-report/guinea/en/.

X. Xu, T. Kypraios, and P. D. O'Neill. Bayesian non-parametric inference for stochastic epidemic models using gaussian processes. *Biostatistics*, pages 1–15, 2016. doi: doi: 10.1093/biostatistics/kxw011.

A Appendix

A.1 Heuristic justification of the adjustment on p

Considering Equation (20), we may refer to the following approximation. For a relatively non-informative prior for p, and so also for $\hat{p}$, the posterior density of $\hat{p}$ is that given by the likelihood function $L(\cdot)$. Regarding the posterior distributions of $\hat{p}$ and p as being modelled by the random variables $\hat{P}$ and P respectively, we can define the variable X as

$$X = P - \hat{P} \text{ and let } \sigma^2 = p^2(1-p)/K_r.$$

Treating σ^2 as a constant, we have $X \sim \mathcal{N}(0, \sigma^2)$. Therefore the multiplicative term $\phi(\hat{p} - p)L(\beta, \gamma, \hat{p}; \boldsymbol{s_r}, \boldsymbol{r_r})$, in (20), is the product of the density of X and $\hat{P}$. Hence, integrating over $\hat{p}$ this gives

$$\int_0^1 \phi(\hat{p} - p)L(\beta, \gamma, \hat{p}; \boldsymbol{s_r}, \boldsymbol{r_r})\ d\hat{p}$$

which by definition is the convolution of X and $\hat{P}$ and, since X and $\hat{P}$ are independent, the density of $P = \hat{P} + X$. This result suggests that the means of $\hat{P}$ and P are the same and the variance of $\hat{P}$ should be increased by the variance of X to obtain the variance of P. In practice the quantity $\sigma^2 = p^2(1-p)/K_r$ can be estimated by using any reasonable estimate of p. If we use $\hat{p}$ to estimate σ^2, the uncertainty about p can be estimated by adding to the posterior variance of $\hat{p}$ the quantity $\hat{p}^2(1-\hat{p})/K_r$.

A.2 Pseudo-code for Methods $1-3$

A.2.1 Method 1

MCMC algorithm using Method 1

1. Start with initial values for β, γ, p and $\boldsymbol{s_r}$.
2. Update the infection times of the reported cases by selecting one individual at random, propose an infection time based on the distribution of the infectious period and accept it with probability A in (19), given the parameters β, γ and p.
3. Update the model parameters:
   (i) Update β according to its full posterior conditional distribution in (17), using the current values of p, γ and $\boldsymbol{s_r}$.
   (ii) Update γ using its full posterior conditional distribution in (18) and the current values of p, β and $\boldsymbol{s_r}$.

(iii) Update p using random walk with acceptance probability

$$A_p = \min\left(\frac{L(\beta, \gamma, p^{new};\, \boldsymbol{s_r}, \boldsymbol{r_r})}{L(\beta, \gamma, p^{old};\, \boldsymbol{s_r}, \boldsymbol{r_r})}, 1\right)$$

using current γ, β and $\boldsymbol{s_r}$.

4. Steps 2 and 3 are repeated until convergence.

A.2.2 Method 2

Algorithm for correction on p

1. Start with initial values for β, γ, $\hat{p}$, p and $\boldsymbol{s_r}$.

2. Update the infection times of the reported cases, $\boldsymbol{s_r}$, using Metropolis-Hastings algorithm, given the current values of parameters β, γ, $\hat{p}$ and p.

3. Update the model parameters:

   (i) Update β according to its full posterior conditional distribution in (17) with p replaced by $\hat{p}$, using the current values of p, $\hat{p}$, γ and $\boldsymbol{s_r}$.

   (ii) Update γ using its full posterior conditional distribution in (18) and the current values of p, $\hat{p}$, β and $\boldsymbol{s_r}$.

   (iii) Update $\hat{p}$ following a random walk scheme with acceptance probability

$$Acc_{\hat{p}} = \min\left(\frac{\phi(\hat{p}^{new} - p)L(\beta, \gamma, \hat{p}^{new};\, \boldsymbol{s_r}, \boldsymbol{r_r})}{\phi(\hat{p}^{old} - p)L(\beta, \gamma, \hat{p}^{old};\, \boldsymbol{s_r}, \boldsymbol{r_r})}, 1\right)$$

   using the current values of p, γ, β and $\boldsymbol{s_r}$.

   (iv) Update p using random walk with acceptance probability

$$Acc_p = \min\left(\frac{\pi_p(p^{new})\phi(\hat{p} - p^{new})}{\pi_p(p^{old})\phi(\hat{p} - p^{old})}, 1\right)$$

   using $\hat{p}$, γ, β and $\boldsymbol{s_r}$. Note that this update is actually independent of β, γ and $\boldsymbol{s_r}$.

4. Steps 2 and 3 are repeated until convergence.

A.2.3 Method 3

Algorithm for approximate Gibbs sampling

1. Start with initial values of β, γ, p and $\boldsymbol{s_r}$.

2. Update the infection times of the reported cases, $\boldsymbol{s_r}$, using Metropolis-Hastings algorithm, given the current values of parameters β, γ and p.

3. Update the model parameters:

   (i) Update β according to its full posterior conditional distribution in (17) using the current values of p, γ and $\boldsymbol{s_r}$.

   (ii) Update γ using its full posterior conditional distribution in (18) and the current values of p, β and $\boldsymbol{s_r}$.

   (iii) Solve Equation (13) and sample the final size K from (14) using current β, γ and $\boldsymbol{s_r}$; conditional on the fact that $K \geq K_r$.

   (iv) Sample the probability of reporting p using (23) using the current final size K.

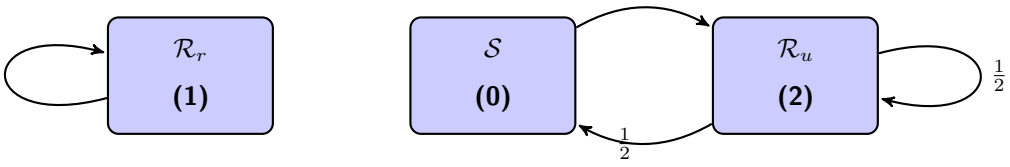
4. Steps 2 and 3 are repeated until convergence.

A.3 Reversible Jump MCMC algorithm

An individual (say k) will always have one of the following states:

- 0 - Susceptible;
  i.e $k \in \mathcal{S}$
- 1 - Removed and reported;
  i.e $k \in \mathcal{R}_r$
- 2 - Removed but not reported;
  i.e $k \in \mathcal{R}_u$.

Below are the possible algorithm transitions presented schematically:



The algorithm is now describe in details as follows:

- Choose an individual at random (let us say k).

- If the state of k is 1 (meaning that the individual was removed and reported), we propose the new infection time s_k using $(s_k - r_k) \sim \text{Exp}(\gamma)$ and s_k is accepted with probability :

$$A_{1\to 1} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})} * \frac{\exp\{-\gamma(r_k - s'_k)\}}{\exp\{-\gamma(r_k - s_k)\}}\right\}$$

where s'_k is the current infection time of the individual k. There is no change in state.

- If the state of k is 0 (susceptible individual) we propose a removal time r_k uniformly in (T_0, T) (the interval (T_0, T) is the time window in which the epidemic has occurred) and an infection time s_k in (T_0, r_k) and add the pair with probability

$$A_{0\to 2} = \min\left\{1, \frac{(T - T_0)(r_k - T_0)}{2} \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r}^{(new)})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r}^{(old)})}\right\}.$$

If this move is accepted, the state of the individual k becomes 2 (which represents individuals that are removed but not reported), i.e

$$|S| = |S| - 1 \qquad \text{and} \qquad |\mathcal{R}_u| = |\mathcal{R}_u| + 1.$$

- If state of k is 2 we either propose, with probability 1/2, to update the pair of infection and removal times, or delete the pair of infection and removal times:

  - Update the pair of infection and removal times of k (with $T_0 \le s_k < r_k \le T$) with probability

$$A_{2\to 2} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})} \frac{r_k - T_0}{r'_k - T_0}\right\}$$

    where r'_k is the removal time of individual k before the new proposed one r_k. The state remains the same.

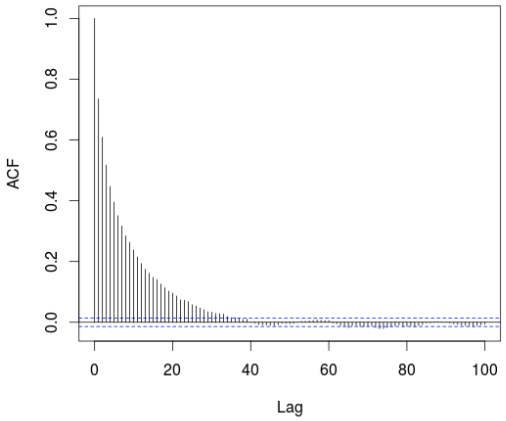
  - Delete the pair of infection and removal times with probability

$$A_{2\to 0} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})} \frac{2}{(T - T_0)(r_k - T_0)}\right\}.$$

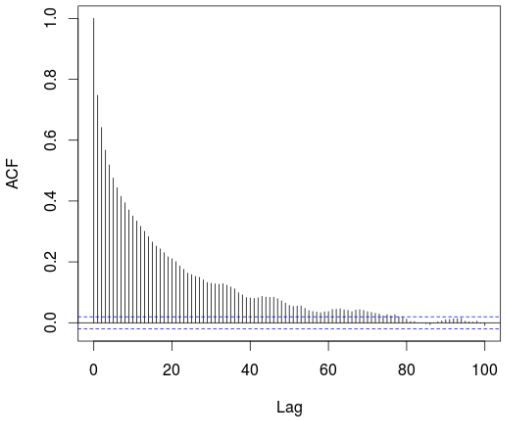
    The state of the individual k becomes 0 if the deletion is accepted, i.e

$$|\mathcal{R}_u| = |\mathcal{R}_u| - 1 \qquad \text{and} \qquad |\mathcal{S}| = |\mathcal{S}| + 1.$$

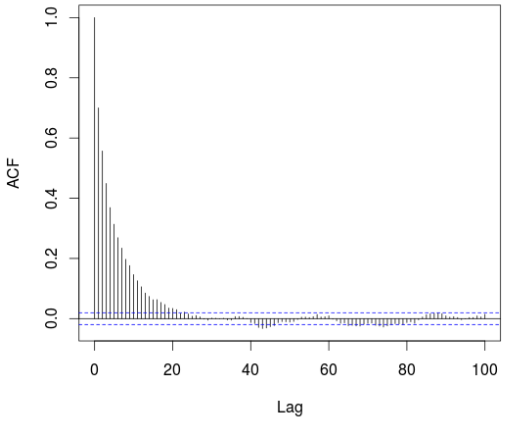
A.4 Auto-correlation functions



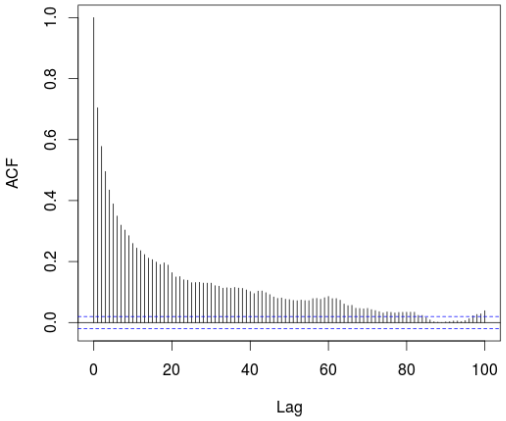
a) ACF plot for γ, Method 1



b) ACF plot for γ, Method 2

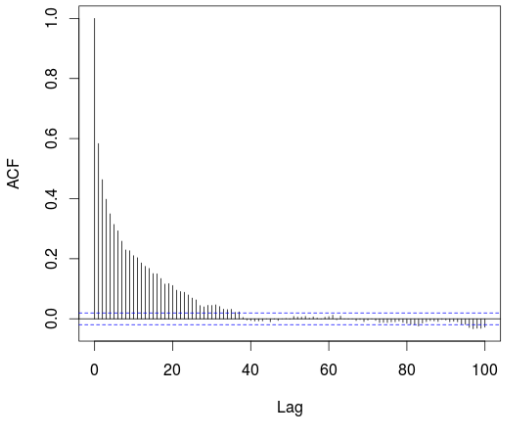


c) ACF plot for γ, Method 3

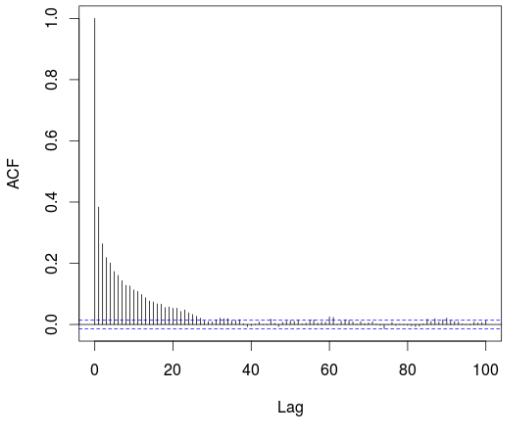


d) ACF plot for γ, Method 4

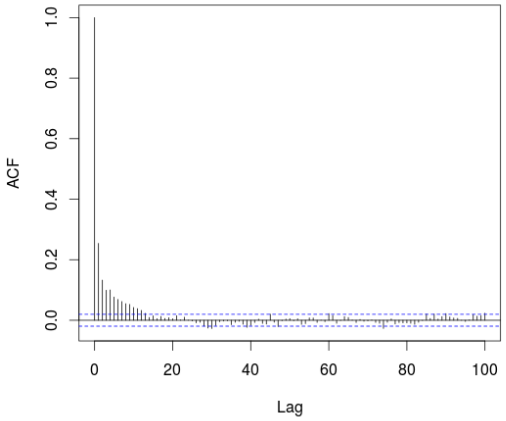
Figure 5: Auto-correlation function for parameter γ (data from Section 4)



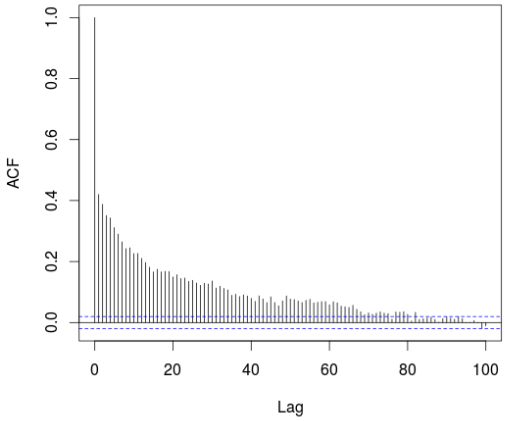
a) ACF plot for K, Method 1



b) ACF plot for K, Method 2



c) ACF plot for K, Method 3



d) ACF plot for K, Method 4

Figure 6: Auto-correlation function for parameter K (data from Section 4)