Computational Gains Via a Discretization of the Parameter
Space in Individual Level Models of Infectious Disease

by
Xuan Fang

A Thesis
Presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada
©Xuan Fang, January, 2012
ABSTRACT

Computational Gains Via a Discretization of the Parameter Space in Individual Level Models of Infectious Disease

Xuan Fang
University of Guelph, 2012

Advisors:
Professors R.Deardon

The Bayesian Markov Chain Monte Carlo (MCMC) approach to inference is commonly used to estimate the parameters in spatial infectious disease models. However, such MCMC analyses can pose a hefty computational burden. Here we present a new method to reduce the computing time cost in such MCMC analyses and study its usefulness. This method is based around the discretization of the spatial parameters in the infectious disease model. A normal approximation of the posterior densities of the original model will be compared to such an approximation of the modified model, using the Kullback-Leibler (KL) divergence measure.
## Contents

I Introduction 1

II Computational Gains Via a Discretization of the Parameter Space in Individual Level Models of Infectious Disease 2

1 Introduction 2

2 Model framework 4

3 MCMC computation and KL-divergence 6

3.1 MCMC analysis for ILM with continuous spatial parameter 6

3.2 MCMC analysis for ILM with a discrete spatial parameter ($\tilde{\beta}$) 8

3.3 Kullback-Leibler divergence (KL-divergence) 9

4 Simulation study 10

4.1 Observed data 11

4.2 Model fitting 11

4.2.1 $P_{(i,t)}$ Model (Continuous $\beta$) 11

4.2.2 $\tilde{P}_{(i,t)}$ Model (Discrete $\tilde{\beta}$) 12
<table>
<thead>
<tr>
<th>4.2.3</th>
<th>Fixed center point of $\beta$ with different count and gap</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.4</td>
<td>Misspecifying $c$</td>
<td>17</td>
</tr>
<tr>
<td>4.3</td>
<td>Computational efficiency</td>
<td>20</td>
</tr>
</tbody>
</table>

5 Discussion and future work 22
Part I

Introduction

Infectious diseases are a major cause of illness and death. Mathematical and statistical models have been introduced in this research area for the purposes of understanding the risk factors associated with disease spread. The individual-level model Deardon et al. [2010] can be used for this purpose, using a Bayesian MCMC approach to fit an ILM to observed data.

However, such an MCMC analysis can be very time consuming, especially if the epidemic data set is large. It would, therefore, be useful if the model could be adjusted in such a way that computation time can be reduced, but without losing desirable epidemic-mosaicing characteristics. An attempt to do this is made in Chapter 2, a manuscript to be submitted to the journal Spatical and Spatio-temporal Epidemiology.
Part II

Computational Gains Via a Discretization of the Parameter Space in Individual Level Models of Infectious Disease

1 Introduction

Infectious diseases can cause severe harm both in terms of public health and economic growth. Typical examples in the public sphere are the SARS epidemic of 2003 (e.g. Aschwanden [2004]) and the H1N1 influenza outbreak of 2009 (e.g. Shan et al. [2010]). The latter resulted in thousands of deaths and billions in costs contributing to a worldwide crisis. Improving warning mechanisms and emergency measures is a challenging task. Much research work has been done in this area [Roux and Aiello, 2005], but much remains to be done.

Mathematical and statistical models can be used to evaluate the potential impact of the transmission of infectious agents [Bailey, 1975]. When attempting to account for the complex spatio-temporal dynamics that many diseases exhibit, standard models such as generalized linear models tend to be inadequate.
The individual-level model of Deardon et al. [2010], are a type of infectious disease model generally applied with the Bayesian framework, with parameters being estimated via Markov Chain Monte Carlo (MCMC) methods. Other recent examples of similar models fitted in such a framework are given by Ster and Ferguson [2007], Jewell et al. [2009], and Cauchemez and Ferguson [2011]. All of these models detail the use of individual-level covariates, such as spatial location, within the model. However, in all cases, the MCMC analysis to be carried out can take a great deal of time, especially for large data sets.

Innovative methods can be used to speed up the MCMC process. The goal of this paper is to introduce a method of changing the parameter space of a standard spatial ILM to increase the speed in which the calculation of the likelihood function can be carried out each MCMC iteration. This new parameterization will be tested via simulation study. Performance of the new model framework will be assessed using the KL-divergence criterion.

The model framework and the MCMC algorithm, as well as an overview of KL-divergence, will be discussed in Sections 2 and 3. The simulation study, and results thereof, will be detailed in Section 4. Finally, a discussion of the results, and some potential avenues for further work are presented in Section 5.
2 Model framework

Here we use the compartmental individual-level model (ILM) framework described in Deardon et al. [2010]. In that framework, a population of n individuals, observed at discrete time points \( t = 1, \ldots, t_{\text{max}} \), is assumed. At each point in time, each individual can be in one of four states: \( S \), \( E \), \( I \) or \( R \). The states are defined as follows:

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S )</td>
<td>Susceptible ( i ) is free of disease and can be infected.</td>
</tr>
<tr>
<td>( E )</td>
<td>Exposed ( i ) is infected, but cannot yet infect others.</td>
</tr>
<tr>
<td>( I )</td>
<td>Infectious ( i ) is infected, and can infect others.</td>
</tr>
<tr>
<td>( R )</td>
<td>Removed ( i ) is removed and cannot now infect others. (e.g. ( i ) has been quarantined, died, or has recovered with acquired immunity).</td>
</tr>
</tbody>
</table>

In this paper, we simplify the compartmental framework to one of only three states: \( S \), \( I \) and \( R \), creating an SIR individual-level model. Each infected individual will respectively move through the states from \( S \) to \( I \) to \( R \). Generally, the initial state of an individual is \( S \). At time point \( t \), each susceptible \( i \) has a probability of being infected by infectious individuals in
the population that is given by:

\[ P_{i,t} = 1 - \exp \left\{ -\alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \right\}, \quad \alpha, \beta \in \mathbb{R}^+ \quad (1) \]

where \( P_{i,t} \) is the probability that \( i \in S(t) \) appears in set \( I(t+1) \), \( \alpha \) is the susceptibility parameter, \( \beta \) is a spatial decay parameter that characterizes the risk of infection over distance, and \( I(t) \) is the set of infectious individuals at time \( t \). Further, \( k(i,j) = d_{ij}^{-\beta} \) is known as the infection kernel, and \( d_{ij} \) is the Euclidean distance between two individuals \( i \) and \( j \).

For \( t = 1, \ldots, t_{\text{max}} \), the likelihood is given by:

\[ L(D|\theta) = \prod_{t=1}^{t_{\text{max}}} \left[ \prod_{i \in S(t+1)} 1 - P_{i,t} \right] \left[ \prod_{i \in I(t+1) \setminus I(t)} P_{i,t} \right] \quad (2) \]

where \( D \) represents the observed data, \( \theta \) is the parameter vector \( \theta = (\alpha, \beta) \), and \( i \in I(t+1) \setminus I(t) \) is the set of individuals that move from \( S(t) \) to \( I(t+1) \). The likelihood function therefore represents the probability of observing all susceptible individuals (which have not been infected) at time \( t+1 \), and newly infectious individuals at time \( t+1 \), multiplied over \( t = 1, \ldots, t_{\text{max}} \).

According to Bayes’ rule, the posterior density is given by:

\[ \pi(\theta|D) = \frac{L(D|\theta)p(\theta)}{\pi(D)} \quad (3) \]
where $\pi(D) = \int \pi(\theta|D)p(\theta)d\theta$ is a normalization constant, and $p(\theta)$ is the prior density of $\theta$. Since $\pi(D)$ cannot usually be calculated, posterior inference is often carried out using simulation-based methods (e.g. MCMC; see Section 3).

Here, all individuals move from $I \rightarrow R$ after an infectious period $\gamma_i$. Further, we use assume $\gamma_i = \gamma$, which is a known scalar, although $\gamma_i$ can easily be estimated as random effects drawn from some distribution.

# 3 MCMC computation and KL-divergence

## 3.1 MCMC analysis for ILM with continuous spatial parameter

MCMC methods are a class of algorithms that are used for sampling from probability distributions via a Markov chain. Under certain conditions, a Markov chain can be constructed to converge to a stationary distribution that is equivalent to the posterior distribution. Hence, output from such a stationary distribution can be treated as a dependent sample from the posterior distribution. Here we use a single-parameter-update random walk Metropolis-Hastings algorithm to achieve this construction. The state at iteration $i$ is denoted, $\theta^{(i)} = (\alpha^{(i)}, \beta^{(i)})$. The algorithm steps are as follows:

Initialization: Set a starting point $\alpha = \alpha^{(0)}, \beta = \beta^{(0)}$. 
Step 1: Generate $Z_\alpha \sim A, Z_\beta \sim B$, where $A$ and $B$ are symmetric distributions centered on zero. Here we use uniform distributions, i.e., $Z_\alpha \sim U[-a, a], Z_\beta \sim U[-b, b], (a, b \in \mathbb{R}^+)$.

Step 2: Propose a new parameter value for $\alpha : \alpha' = \alpha^i + Z_\alpha$.

Step 3: Calculate the acceptance probability $\psi_\alpha$ as:

$$
\psi_\alpha = \min \left(1, \frac{\pi(\alpha', \beta^{(i)} | D)}{\pi(\alpha^i, \beta^{(i)} | D)} \right),
$$

Step 4: Accept $\alpha^{(i+1)} = \alpha'$ with probability $\psi_\alpha$, or else reject and set $\alpha^{(i+1)} = \alpha^{(i)}$.

Step 5: Propose a new parameter value for $\beta : \beta' = \beta^i + Z_\beta$.

Step 6: Calculate the acceptance probability $\psi_\beta$ as:

$$
\psi_\beta = \min \left(1, \frac{\pi(\alpha^{(i+1)}, \beta' | D)}{\pi(\alpha^{(i+1)}, \beta^{(i)} | D)} \right),
$$

Step 7: Accept $\beta^{(i+1)} = \beta'$ with probability $\psi_\beta$, or else reject and set $\beta^{(i+1)} = \beta^{(i)}$.

Step 8: Go back to Step 1.

We typically need to recalculate the posterior thousands of times for different values of the parameters to effectively sample from the posterior distribution via MCMC. Therefore, for large data sets, for which the likelihood may take a long time to compute, it may require a long time to carry out an MCMC analysis.
3.2 MCMC analysis for ILM with a discrete spatial parameter ($\tilde{\beta}$)

Recall the previous model in Section 2. We now consider a discretization of the $\beta$ parameter, giving the new model

$$\tilde{P}_{i,t} = 1 - \exp \left\{ -\alpha \sum_{j \in I(t)} d_{ij}^{-\tilde{\beta}} \right\}, \alpha \in \mathbb{R}^+, \tilde{\beta} \in \{b_1, \ldots, b_n\}$$

$$b_i \in \mathbb{R}^+, i = 1, \ldots, n$$

(4)

When using MCMC to fit this model, we replace Step 5 of the MCMC algorithm in Section 3.1 with:

Step 5: Randomly select proposed $\tilde{\beta}, \tilde{\beta}(i)$ from $\{b_1, \ldots, b_n\}$.

We now explain how such a discretization can lead to a saving in computation time. Recall the likelihood function of (2) and take logs to get the log-likelihood function:

$$\log L(D|\theta) = \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \log (1 - \tilde{P}_{i,t}) + \sum_{t=1}^{t_{\text{max}}} \sum_{i \in I(t+1) \setminus I(t)} \log \tilde{P}_{i,t}$$

(5)

Now, consider the following part of this function:
\[
\sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \log (1 - \tilde{P}_{i,t}) = \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \log \left[ \exp \left( -\alpha \sum_{j \in I(t)} d_{ij}^{-\tilde{\beta}} \right) \right] \\
= \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \left( -\alpha \sum_{j \in I(t)} d_{ij}^{-\tilde{\beta}} \right) \\
= -\alpha \psi(\tilde{\beta}) 
\]

where

\[
\psi(\tilde{\beta}) = \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \sum_{j \in I(t)} d_{ij}^{-\tilde{\beta}} 
\]

As \( \tilde{\beta} \) can only take a finite set of values, \( \psi(\tilde{\beta}) \) can be calculated for each of those fixed values of \( \tilde{\beta} \) before the MCMC algorithm is run. Each time we calculate the (log)likelihood function, according to the current \( \tilde{\beta} \) value, one of the \( \psi(\tilde{\beta}) \) values can be called directly to replace a major part of the likelihood calculation. In this way, computing time can be substantially reduced. Of course, the downside to this is that the pre-defined set \( \{b_1, \ldots, b_n\} \) needs to be chosen with care.

### 3.3 Kullback-Leibler divergence (KL-divergence)

The KL divergence is a non-symmetric measure of the difference between two probability densities. Considering two den-
sity distributions $g(x)$ and $h(x)$ for $x \in D$, the value of the KL-divergence between them is given by:

$$d(g, h) = \int_{x \in D} g(x) \log \frac{g(x)}{h(x)} dx$$

(8)

To calculate (8), $g(x)$ and $h(x)$ are required to be normalized densities. This makes it problematic to calculate the KL-divergence for posteriors which are estimated via Monte Carlo methods. However, if $g(x)$ and $h(x)$ are the probability density functions of two Gaussian distributions, $g(x) = N(x; \mu_q, \sigma_q^2)$ and $h(x) = N(x; \mu_p, \sigma_p^2)$, the KL-divergence is given by:

$$KL_{N_1}(\mu_q, \sigma_q; \mu_p, \sigma_p) = 0.5 \log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 + \sigma_q^2 - 2\mu_q\mu_p}{2\sigma_p^2} - 0.5$$

(9)

See Penny [2001] for further details.

4 Simulation study

Results of the simulation study are considered here. Three key factors which describe the value of $\tilde{\beta}$ chosen in the discrete analysis are $n$ (the number of $\tilde{\beta}$ values), $c$ (the centre point of the range of $\tilde{\beta}$ values) and $\Delta$ (the gap between adjacent pairs of $\tilde{\beta}$ values). We use the following notation to define the $\tilde{\beta}$ parameter space. The set of values $\tilde{\beta}$ can take is given by:
\[ C_\beta = \left\{ c - \frac{(n-1)}{2} \Delta, \ldots, c - 2\Delta, c - \Delta, c, c + \Delta, c + 2\Delta, \ldots, c + \frac{(n-1)}{2} \Delta \right\} \]

Note that here, only odd values of \( n \) are considered.

4.1 Observed data

The population here consists of 625 individual points on a 25×25 grid area. That is, coordinates of the individuals are given by all combinations \((x, y)\) for \( x, y = 1, \ldots, 25 \). When \( t = 1 \), one individual point is randomly picked as the source of the infectious disease outbreak. Here we set \( t_{\text{max}} = 15 \), \( \alpha = 0.5 \), \( \beta = 3.5 \) and \( \gamma_i = 4 \). The epidemic is then simulated via (1) in Section 2 and continues until \( t = t_{\text{max}} \).

4.2 Model fitting

4.2.1 \( P_{(i,t)} \) Model(Continuous \( \beta \))

Using the data from the epidemic simulation above, the model in Section 2 is fitted via MCMC with a starting point of \( \alpha = 3 \) and \( \beta = 5 \). A random-walk Metropolis-Hastings algorithm is run for 3000 iterations with a burn-in of 150 iterations being removed. Figure 1 shows a plot of a typical marginal posterior distribution of \( \beta \), along with a normal distribution with the mean and variance set equal to the posterior mean and variance respectively. We can see the approximation of the normal to the marginal posterior distribution is reasonable. Of course, normal
Figure 1: Marginal posterior density of $\beta$ values (the solid line) from MCMC outputs and normal approximation of $\beta$ (the broken line).

distribution can also be used to approximate discrete distribution cases (Quinn and MacGillivray [1986] and Govindarajulu [1965]).

4.2.2 $\tilde{P}_{(i,t)}$ Model(Discrete $\tilde{\beta}$ )

Here we use an independence sample Metropolis-Hastings step to update the $\tilde{\beta}$ parameter as described in Section 3. That is,
newly proposed values of $\tilde{\beta}$ are selected randomly from a set $\tilde{\beta}' = b_1, ..., b_n$.

Here we approximate the marginal posterior of $\tilde{\beta}$ using a normal distribution with mean and variance equal to the estimate posterior mean and variance. This conveniently allows us to compare the normally approximated posterior of $\beta$ under the $P_{(i,t)}$ model and the $\tilde{\beta}$ under the $\tilde{P}_{(i,t)}$ using the KL-divergence measure for two normal distributions, as discussed in Section 3.2.

### 4.2.3 Fixed center point of $\beta$ with different count and gap

Here $c$ is set equal to the true $\beta$ used in the epidemic simulation described in Section 4.2.1 ($c = 3.55$). We calculate the KL-divergence for the normal approximation of the marginal posterior density of $\beta$ under the $P_{(i,t)}$ model and the $\tilde{\beta}$ under $\tilde{P}_{(i,t)}$ model for various values of $n$ and $\Delta$. To show the replicability of the results, 10 replicated epidemics are generated and used here. Table 1 shows the mean KL-divergence and the standard error of these 10 replications. We can see that in all cases, as $n$ increases, the KL-divergence decreases. In all cases, as $\Delta$ increases, the KL-divergence decreases.

Figure 2 shows the normally approximated posterior for the combinations of $(c, \Delta, n)$ for one typical realization. It is clearly evident that similarity between the two approximate posteriors
Figure 2: Normal approximation of marginal posterior density of $\beta$ under the $P_{(i,t)}$ model (solid line) and the $\tilde{\beta}$ under $\tilde{P}_{(i,t)}$ model (dotted line). $n \in \{3, 5, 7, 9\}$, $\Delta \in \{0.03, 0.05, 0.07\}$, $c = 3.55$. 
Table 1: Mean KL-divergence (and standard error) for ten simulated epidemics: \( n \in \{3, 5, 7, 9\} \), \( \Delta \in \{0.03, 0.05, 0.07\} \), \( c = 3.55 \).

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>0.03</th>
<th>0.05</th>
<th>0.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.3555</td>
<td>1.3596</td>
<td>0.5049</td>
</tr>
<tr>
<td></td>
<td>(0.0579)</td>
<td>(0.0442)</td>
<td>(0.0243)</td>
</tr>
<tr>
<td>5</td>
<td>1.9249</td>
<td>0.2459</td>
<td>0.0416</td>
</tr>
<tr>
<td></td>
<td>(0.0513)</td>
<td>(0.0216)</td>
<td>(0.0037)</td>
</tr>
<tr>
<td>7</td>
<td>0.4622</td>
<td>0.0481</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>(0.0238)</td>
<td>(0.0055)</td>
<td>(0.0043)</td>
</tr>
<tr>
<td>9</td>
<td>0.1731</td>
<td>0.0112</td>
<td>0.0067</td>
</tr>
<tr>
<td></td>
<td>(0.0143)</td>
<td>(0.0043)</td>
<td>(0.0028)</td>
</tr>
</tbody>
</table>

increases with increasing \( n \), and with increasing \( \Delta \).

Now we consider large values of \( \Delta \). We calculate the KL-divergence between the normal approximations of marginal posterior density of \( \beta \) under the \( P_{(i,t)} \) model and the \( \tilde{\beta} \) under the \( \tilde{P}_{(i,t)} \) model where \( \Delta \in \{0.1, 0.15, 0.2, 0.25, 0.3\} \), \( n \in \{3, 5\} \), \( c = 3.55 \). Once again, results are averaged over 10 simulated epidemics. Table 2 shows the mean KL-divergence and standard errors for those. Figure 3 contains plots for the cases \( n \in \{3, 5\} \) and \( \Delta \in \{0.1, 0.15, 0.2, 0.25, 0.3\} \) (for one typical realization). It can be seen, for both \( n = 3 \) and \( n = 5 \), that values of \( \Delta \) that are too low result in relatively high KL-divergence, and, similarly, values of \( \Delta \) that are too high result in relatively high KL-divergence.
Figure 3: Normal approximation of marginal posterior density of $\beta$ under the $P(i,t)$ model (solid line) and the $\tilde{\beta}$ under $\tilde{P}(i,t)$ model (dotted line). $n \in \{3, 5\}$, $\Delta \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$, $c = 3.55$. 

16
Table 2: Mean KL-divergence (and standard error) for ten simulated epidemics: $n \in \{3, 5\}, \Delta \in \{0.1, 0.15, 0.2, 0.25, 0.3\}, c = 3.55$.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1004</td>
<td>0.0133</td>
<td>0.0239</td>
<td>0.0991</td>
<td>1.1129</td>
</tr>
<tr>
<td></td>
<td>(0.0111)</td>
<td>(0.0022)</td>
<td>(0.0076)</td>
<td>(0.0078)</td>
<td>(0.0638)</td>
</tr>
<tr>
<td>5</td>
<td>0.0091</td>
<td>0.0053</td>
<td>0.0439</td>
<td>0.2672</td>
<td>0.2939</td>
</tr>
<tr>
<td></td>
<td>(0.0015)</td>
<td>(0.0009)</td>
<td>(0.0206)</td>
<td>(0.0371)</td>
<td>(0.0451)</td>
</tr>
</tbody>
</table>

4.2.4 Misspecifying $c$

Of course, in reality we will not know the true $\beta$ used to “generate” the epidemic. Therefore here, we fit the $\tilde{P}_{(i,t)}$ model and examine results for $c \neq 3.55$. Figure 4 shows the marginal posterior mass functions of $\tilde{\beta}$ for typical epidemic runs under the $\tilde{P}_{(i,t)}$ model, where $n = 7, c = 3.3, \Delta \in \{0.02, 0.05, 0.1, 0.15\}$.

When $\Delta$ is small and $c$ is far from the true value, we can see we get a heavily skewed mass function for which the mode equals to $\max(C_\beta)$. This would mean that the normal approximation would not be a good one here. More importantly it is an indication that our range of $\tilde{\beta}$ is not wide enough and implies we need to widen it. To confirm this, we increase $\Delta$ and see that when $\Delta$ is large enough, the output will be for more approximately normal.

Figure 5 shows normally approximated posteriors for $\Delta \in \{0.02, 0.05, 0.1, 0.15\}, c \in \{3.3, 3.4, 3.5, 3.6, 3.7, 3.8\}$ and $n = 7$. We see that, if the $c$ value is far from the true value then $\Delta$
Figure 4: Histograms of the marginal density of the $\tilde{\beta}$ under $\tilde{P}_{(i,t)}$ model, $n = 7$, $c = 3.3$, $\delta \in \{0.02, 0.05, 0.1, 0.15\}$.
Figure 5: Marginal posterior density of $\beta$ under the $P_{(i,t)}$ model (solid line “true”) and the $\tilde{\beta}$ under $\tilde{P}_{(i,t)}$ model (dotted lines). $n = 7$, $\Delta \in \{0.02, 0.05, 0.1, 0.15\}$, $c \in \{3.3, 3.4, 3.5, 3.6, 3.7, 3.8\}$. 

19
needs to be relatively large in order to achieve acceptable performance.

Figure 6 shows the marginal posterior, and respective normal approximated posteriors, distributions for situations in which $c \ll 3.55$. Specifically, we set $c = 1.5$, $n \in \{31, 39, 51\}$, and $\Delta = 0.1$. The output makes no sense until the set of $\tilde{\beta}$ has an intersection with broad range of $\beta$ values for which there is high posterior mass under the $P_{(i,t)}$ model. That can be achieved by ensuring a large enough value of $n$ is chosen, even around a heavily misspecified value of $c$.

### 4.3 Computational efficiency

Tests show that when the new method is used, the computing time for the whole MCMC process can be reduced significantly. This test is run on a personal computer with an Intel Core i5 750 CPU running at 2.67 GHz with 3GB memory. Table 3 shows that almost 75% of computing time is saved when five different $\beta$ values are used in the discrete $\tilde{\beta}$ model for a typical epidemic replication. We can also see that when $n$ or $\Delta$ changes, the percentage of time saving does not significantly differ.

In situations where the epidemic strength is similar, but the population is large, the time saving due to discretization might be expected to be even greater. This is because, the epidemic would likely last longer, and more effort would need to be
Figure 6: Histograms and the normal approximation of marginal posterior of density of the $\tilde{\beta}$ under $\tilde{P}(i,t)$ model, $c = 1.5$, $\Delta = 0.1$, $n \in \{31, 39, 51\}$.

<table>
<thead>
<tr>
<th>$P(i,t)$ model</th>
<th>$\tilde{P}(i,t)$ model</th>
<th>$\tilde{P}(i,t)$ model</th>
<th>$\tilde{P}(i,t)$ model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 5$</td>
<td>$n = 5$</td>
<td>$n = 7$</td>
<td>$n = 5$</td>
</tr>
<tr>
<td>$\Delta = 0.05$</td>
<td>$\Delta = 0.05$</td>
<td>$\Delta = 0.05$</td>
<td>$\Delta = 0.1$</td>
</tr>
<tr>
<td>Time cost</td>
<td>836</td>
<td>206</td>
<td>207</td>
</tr>
</tbody>
</table>

Table 3: Computation time under different models with varying count and gap
spent calculating \( \psi(\beta) = \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \sum_{j \in I(t)} d_{ij}^{-\beta} \) under the \( P_{(i,t)} \) model (see Section 3.2). A weaker epidemic, that does not die out quickly, might also lead to data for which the effects of discretization would be highly beneficial, for the same reason. Of course, a possible computational downside to using the \( \tilde{P}_{(i,t)} \) model would be exhibited if many MCMC analyses were required in order to desire a suitable set of \( \tilde{\beta} \) to be used. Similarly, if a large number of \( \tilde{\beta} \) are chosen (perhaps to avoid this “\( \tilde{\beta} \) tuning” problem) the initial calculation of \( \psi(\tilde{\beta}) \) could take a very long time. However, unless very extreme numbers of \( \tilde{\beta} \) are chosen, this would likely be mitigated by the time saving during each MCMC iteration.

5 Discussion and future work

The main purpose of this paper is to find an effective method to reduce computation time when individual-level models are fitted via MCMC. Obviously, in the real world situation, quick and accurate reporting is highly desirable so that modeling-deduced conclusions might be brought to bear on policy in a timely and productive manner (for example, for constructing control strategies for foot and mouth disease).

The simulation study in Section 4.2 shows it is possible to discretize the spatial parameter and achieve results consistent
with the original continuous parameter space. This effect can be achieved by careful selection of the values of $\tilde{\beta}$ considered, itself through careful selection of $\Delta$, $n$, and $c$. So the remaining problem becomes how to accomplish this selection. In general, increasing the number of discrete $\tilde{\beta}$ values could have a positive effect on the final posterior result, but it obviously increases the computation time. How can we balance these effects? Table 2 shows that increasing $\Delta$ can have both negative and positive effects on simulation.

The time spent determining suitable values of $\tilde{\beta}$ is also a part of the total analysis time. We show in Section 4.3 that the new method can save around 75% of computation time over the old model. This would suggest that, if the time cost of determining acceptable values of $\tilde{\beta}$ can be controlled to below 75% of the analysis time associated with the original model, then the discretization of $\tilde{\beta}$ is a valuable endeavor. If the time required to determine suitable values of $\tilde{\beta}$ is too great, there may be no gain (and indeed serious loss) in using the new model. Further research conducted in this area would certainly appear to be warranted.

A number of possible approaches suggest themselves. One is to choose a liberal set of $\tilde{\beta}$; meaning a very large number of $\tilde{\beta}$, spread over a large range of the original potential $\Delta$-space. Another is to take a sequential approach whereby a small num-
ber of $\tilde{\beta}$ are used over a relative spall interval of $\beta$ space, and then results from the analysis are used to adapt the $\tilde{\beta}$ used in a subsequent phase. For example, if the posterior mode for the discrete set of $\tilde{\beta}$ is not the minimum or maximum value of $\tilde{\beta}$, that would offer some evidence about the location of the true model and the extent of the posterior dispersion around this mode. Alternatively, if the posterior mode is found to be the minimum (or maximum) of the $\tilde{\beta}$ values used, this would imply that the time posterior mode is lower than the values so far used. Each eventually could be used to help devise a new set of $\tilde{\beta}$ to be considered. Various steepest-ascent or imputation methods could be used to derive this new set from the current information obtained.

Another idea might be to partition the data in some way such that the full model can be fitted to a subset of the data; a small temporal window might be sensible. From this analysis a posterior over continuous $\beta$ space could be obtained which could indicate a set of sensible $\tilde{\beta}$ to be used in an MCMC analysis of the full data set.

An obvious downside to the use of the $\beta$-discretization method of this paper is its reliance on certainty in the data. In infectious disease modeling, vital data such as the times of infection are generally impossible to obtain with certainty. In such situations data augmented MCMC is often used to incorporate such
uncertainty. Unfortunately, the method used here would be of little use in such a situation. It is therefore suggested that time-saving methods, such presented here, only be used for very large data sets in which there might be little choice but to find some computational trick to make practicable inference possible.

References


Z. Govindarajulu. Normal approximations to the classical dis-


