Modelling and Bayesian Inference for Structured-population Epidemics

Philip D. O'Neill, University of Nottingham

1. Introduction

2. Temporal: NLV outbreak

3. Non-temporal: Two-level mixing model

1. Introduction

Modelling

We consider stochastic epidemic models with parameters θ.

Parameters typically govern
- infection mechanism;
- infectious period;
- vaccination status;
- population heterogeneity;
- etc...

Data

Typical data, X, are
- final outcome data;
- incomplete temporal data.

Data may include covariates,
eg age, location, etc.

Objective

Inference about θ given data X.

In a Bayesian framework the objective requires calculation of the posterior density

$$\pi(\theta|X) = \frac{\pi(X|\theta)\pi(\theta)}{\int_\Theta \pi(X|\theta)\pi(\theta)d\theta},$$

i.e.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Problem:
The normalising integral is typically intractable.

Solution:
If we can generate random samples from the posterior distribution, then the distribution (or any summary statistics) can be estimated.

We shall use Markov Chain Monte Carlo (MCMC) to generate approximate samples from $\pi(\theta|X)$.

What is MCMC?

Aim is to simulate samples from a density π which we only know up to proportionality.

MCMC works by defining a Markov Chain with π as its stationary distribution; then run chain for a long time, and take samples.

Two common methods are:
(i) Gibbs sampling : Target is $\pi(x_1, \ldots, x_N)$, method works by updating each component x_i according to its marginal conditional distribution:

$$\pi(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N).$$

(ii) Metropolis Hastings : If the chain is at x, propose a new position y according to $q(y|x)$, then accept with probability

$$\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)} \wedge 1.$$

Why MCMC?

Flexibility - existing methods often rely on using unrealistic simplifying model assumptions

Good for missing data problems

Implementation - can be straightforward (but not always)

Naturally allows Bayesian framework

2. Temporal data: NLV outbreak

(+ PJ Marks, Public Health, Nottingham)

Background

Outbreak of gastroenteritis in summer 2001 at a school in Derbyshire, England.

A single strain of Norwalk-like virus (NLV) was found to be the causative agent.

Believed to be person-to-person spread.

Of 492 children, 186 showed symptoms.

School has 15 classrooms; each child is based in one.

Data

Absence records plus questionnaires.
Include age, period of illness, times of vomitting episodes in classrooms.

Objective

Explore the role of vomiting in infection spread.

Example: Classroom 10

Here, three children vomited in class on day 10.

Ten children were absent/ill the next day (11), and a further two on day 12.

Six children were already absent/ill before the vomiting episode.

General There were 15 children in total involved in vomiting episodes.

Per-classroom attack rates (ignoring temporal data) are higher in vomit-episode classes.

Modelling assumptions

For each day during the outbreak,
each of the 492 pupils is classified as one of

absent
susceptible
infectious
returning
vomiter

using the available data and extra modelling assumptions.

E.g.
first absent on day t (data)
implies
infectious on day $t-1$ (assumption).

Stochastic transmission model

A susceptible on weekday t remains so on day $t+1$ provided they avoid infection from:

each classroom infective with probability q_c;
each school infective with probability q_s;
each classroom vomiter with probability q_v.

Independence between each susceptible-infective pair is assumed.

A susceptible on weekend-day t remains so on day $t+1$ if they avoid infection from each infective with probability q.

For classroom j, $j = 1, \ldots, 15$, and day t, let

S_t^j denote number of susceptibles;
I_t^j denote number of infectives;
V_t^j denote number of vomiters;
R_t^j denote number of returners,

$$S_t = \sum_{j=1}^{15} S_t^j, \quad \text{etc.}$$

Then for $t =$ Tuesday to Saturday,

$$P[S_t^j = k] = \binom{R_{t-1}^j + S_{t-1}^j}{k} p^k (1-p)^{R_{t-1}^j + S_{t-1}^j - k}$$

where

$$p = q_c^{I_{t-1}^j} q_s^{I_{t-1}} q_v^{V_{t-1}^j}$$

while if t is a Sunday or Monday,

$$P[S_t = k] = \binom{R_{t-1} + S_{t-1}}{k} p^k (1-p)^{R_{t-1} + S_{t-1} - k}$$

where

$$p = q^{I_{t-1}}.$$

Role of vomiters

We consider two models:

M_1 has parameters (q_c, q_s, q_v, q)
M_2 has parameters $(q_c, q_s, q_v = q_c, q)$

i.e. M_1 is full model;
M_2 classes vomiters as infectives.

Question:
Which model is more likely given the data?

So we are interested in

$$P(M_i|\text{ data}).$$

MCMC using one model

A simple Metropolis-Hastings algorithm is defined as follows.

Let $\theta = (q_c, q_s, q_v, q)$.

Propose a new value θ^* with density

$$g(\theta^*|\theta)$$

and accept θ^* with probability

$$\frac{\pi(\text{data}|\theta^*)}{\pi(\text{data}|\theta)}\frac{\pi(\theta^*)}{\pi(\theta)}\frac{g(\theta|\theta^*)}{g(\theta^*|\theta)} \wedge 1.$$

This can be done one parameter at a time, or all together.

Reversible Jump MCMC

Can extend standard MCMC to allow parameter-dimension changing moves (Green, 1995).

Within model i ($i = 1, 2$), jump with probability p_i.

Consider a 'dummy' variable u:

$$(q_c^1, q_v^1) \leftrightarrow (q_c^2, u)$$

$M_2 \to M_1$:

Sample u from $N(0, \sigma^2)$, with density $\phi(u)$.

Propose $q_c^1 = q_c^2$; $q_v^1 = q_c^2 + u$.

Accept with probability

$$\frac{\pi(q_c^1, q_v^1)}{\pi(q_c^2)} \frac{p_1}{p_2} \frac{1}{\phi(u)} \wedge 1.$$

Reverse jump is similar, but u is non-random.

Results

Full model only

	q_c	q_s	q_v	q
Mean	0.9976	0.9984	0.9836	0.9995
S.dev.	0.00127	0.000156	0.0108	0.000082

Two-model set-up

$$P(M_1|\text{data}) = 0.04$$

So data do not suggest that $q_v \neq q_c$.

Artificial data

An artificial dataset was created in which **all** new infections in classroom 10 occurred 1 day after the (three-child) vomitting incident. We would therefore expect to see more posterior support for model 1.

Results

$$P(M_1|\text{data}) = 0.18$$

Comments

Work is preliminary

Data quality is less than ideal

Model assumptions could be altered

Methodology appears flexible

Callibration issues - what can be detected?

3. Non-temporal:
Two-level mixing models

(+ N Demiris, Nottingham)

Consider a population of size N, split into groups (eg households).

Assume S-I-R model.

In a two-level mixing household model, each infectious individual with infectious period T_I can:

- infect household members with probability

$$1 - \exp(-\lambda_L T_I);$$

- infect any individual with probability

$$1 - \exp(-\lambda_G T_I / N).$$

Given final numbers $\boldsymbol{x}$ ultimately infected in a population of households, the posterior of interest is

$$\pi(\lambda_L, \lambda_G | \boldsymbol{x}) \propto \pi(\boldsymbol{x} | \lambda_L, \lambda_G) \pi(\lambda_L, \lambda_G).$$

Problem

The likelihood $\pi(\boldsymbol{x}|\lambda_L, \lambda_G)$ is computationally intractable for all but small population sizes.

A solution

Find a latent variable Y such that

$$\pi(\boldsymbol{x}|\lambda_L, \lambda_G, Y)$$

and

$$\pi(Y|\lambda_L, \lambda_G)$$

are both tractable.

Then we can work with the augmented posterior density

$$\pi(\lambda_L, \lambda_G, Y|\boldsymbol{x}) \propto \pi(\boldsymbol{x}|\lambda_L, \lambda_G, Y)\pi(Y|\lambda_L, \lambda_G)\pi(\lambda_L, \lambda_G).$$

Y is final severity

The final severity of an epidemic is

$$T_\infty = \sum_{i=1}^{R_\infty} T_i,$$

where R_∞ is the final number infected.

If N is large then

(i) Given T_∞, the fates of different households are approximately independent (Ball *et al*, 1997). Thus

$$\pi(\boldsymbol{x}|\lambda_L, \lambda_G, T_\infty)$$

can be calculated easily.

Specifically, each individual independently avoids global infection with probability

$$\exp(-\lambda_G T_\infty/N).$$

Conditioning on the number infected globally, it is then straightforward to find the distribution of the total number infected in a household.

(ii) If the epidemic takes off, the quantity

$$\sqrt{m}\left(\frac{T_\infty}{N} - \mu(\lambda_L, \lambda_G)\right),$$

where m denotes the number of groups, converges in distribution to a Gaussian random variable with mean 0 and variance $\sigma^2(\lambda_L, \lambda_G)$.

Here, $\mu(\lambda_L, \lambda_G)$ and $\sigma^2(\lambda_L, \lambda_G)$ are both known quantities.

Threshold Parameter

$$R_* = \lambda_G E[T_I]\nu,$$

where ν is the mean size of a group outbreak, initiated by a randomly chosen individual, in which only local infections are permitted.

Application to influenza data

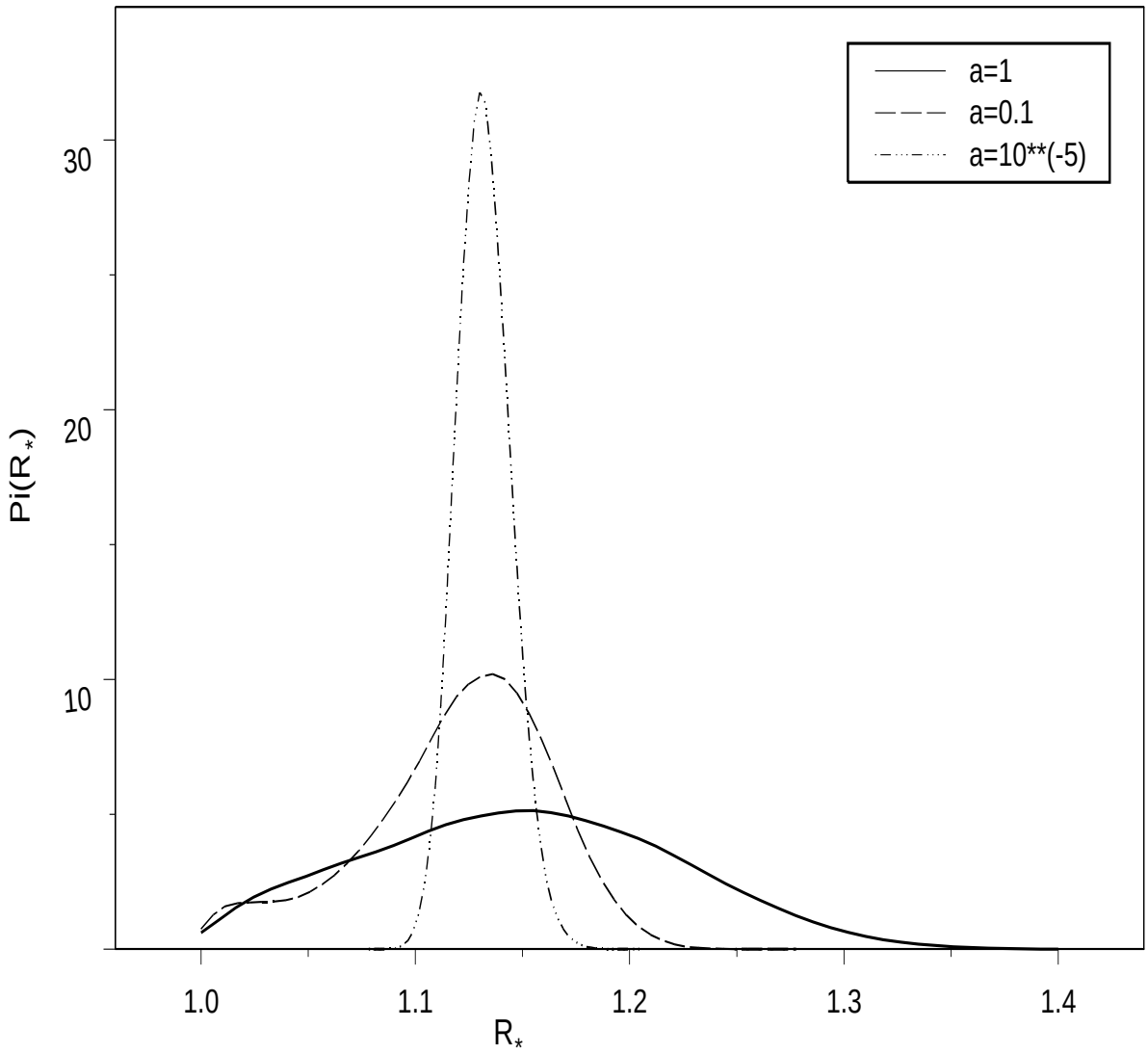
Combined data on influenza outbreaks in Tecumseh, Michigan.

	Susceptibles per household				
No. infected	1	2	3	4	5
0	110	149	72	60	13
1	23	27	23	20	9
2		13	6	16	5
3			7	8	2
4				2	1
5					1
Total	133	189	108	106	31

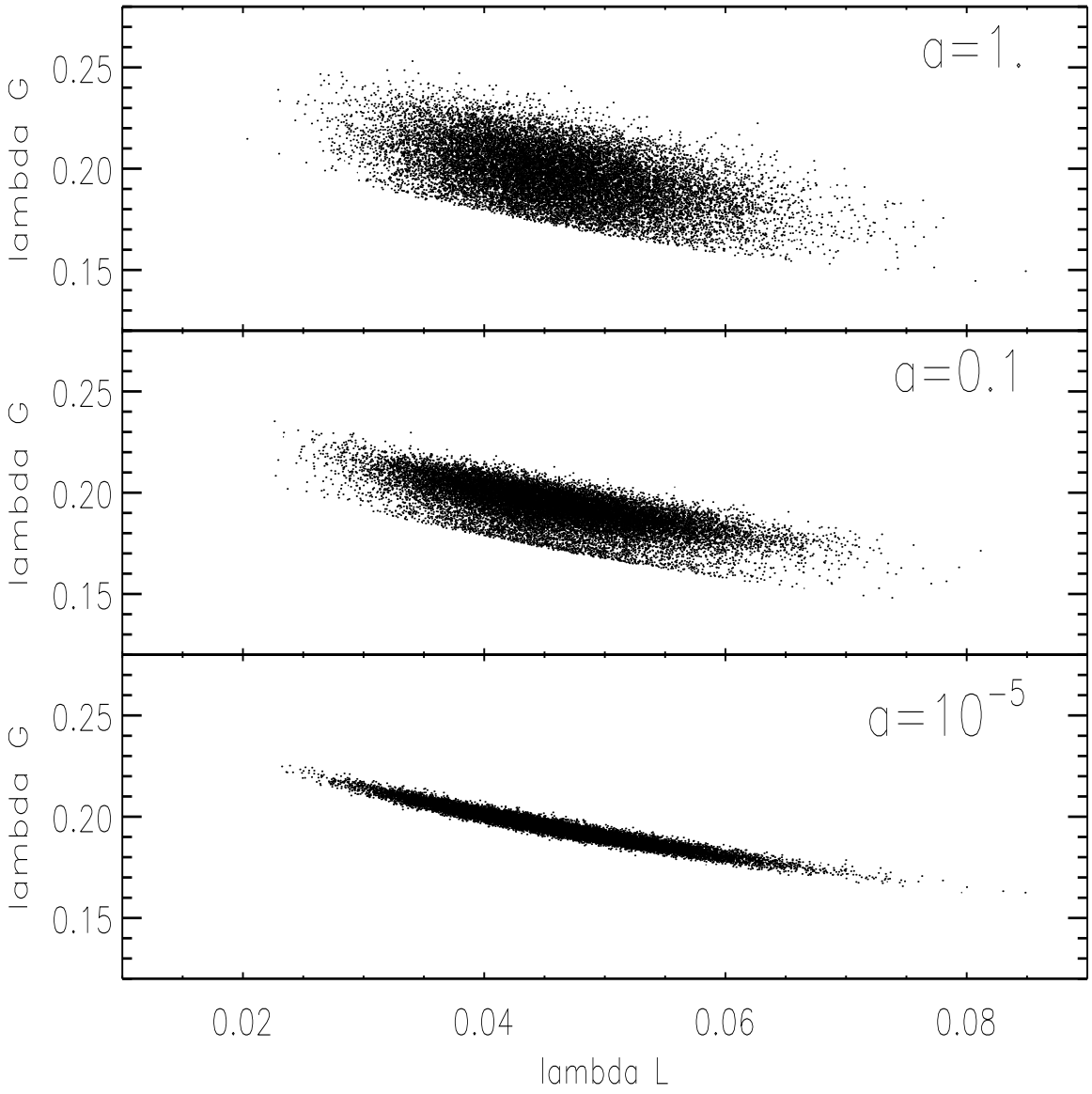
Data are a 10% sample from the population.

Notation: $\alpha = 0.1$ is proportion observed.

Posterior density plots for R_*, assuming $\alpha = 1, 0.1$ and 10^{-5}.



Scatterplot of λ_L and λ_G assuming $\alpha = 1, 0.1$ and 10^{-5}, respectively, from top to bottom.



Investigating the approximation: one-level mixing model

Suppose now that all households are size 1.

Now λ_L is irrelevant and the model is of homogeneous S-I-R type.

Given T_∞, the approximation pseudolikelihood is Binomial(N, p), where

$$p = 1 - \exp(-\lambda_G T_\infty / N).$$

Thus

$$P(R_\infty = x) = \binom{N}{x} p^x (1-p)^{N-x}.$$

However, for this model the final size distribution can be evaluated exactly using a (standard) set of recursive equations.

Results

E.g. $N = 120$, observe $x = 30$ cases.

Exact:
$E(\lambda_G|x) = 0.296$, $var(\lambda_G|x) = 0.00435$

Approximation:
$E(\lambda_G|x) = 0.302$, $var(\lambda_G|x) = 0.000937$

Approximation overestimates, and has less posterior variance.

Probable reason: approximation neglects the probability that epidemic might not take off.

Likelihood comparison

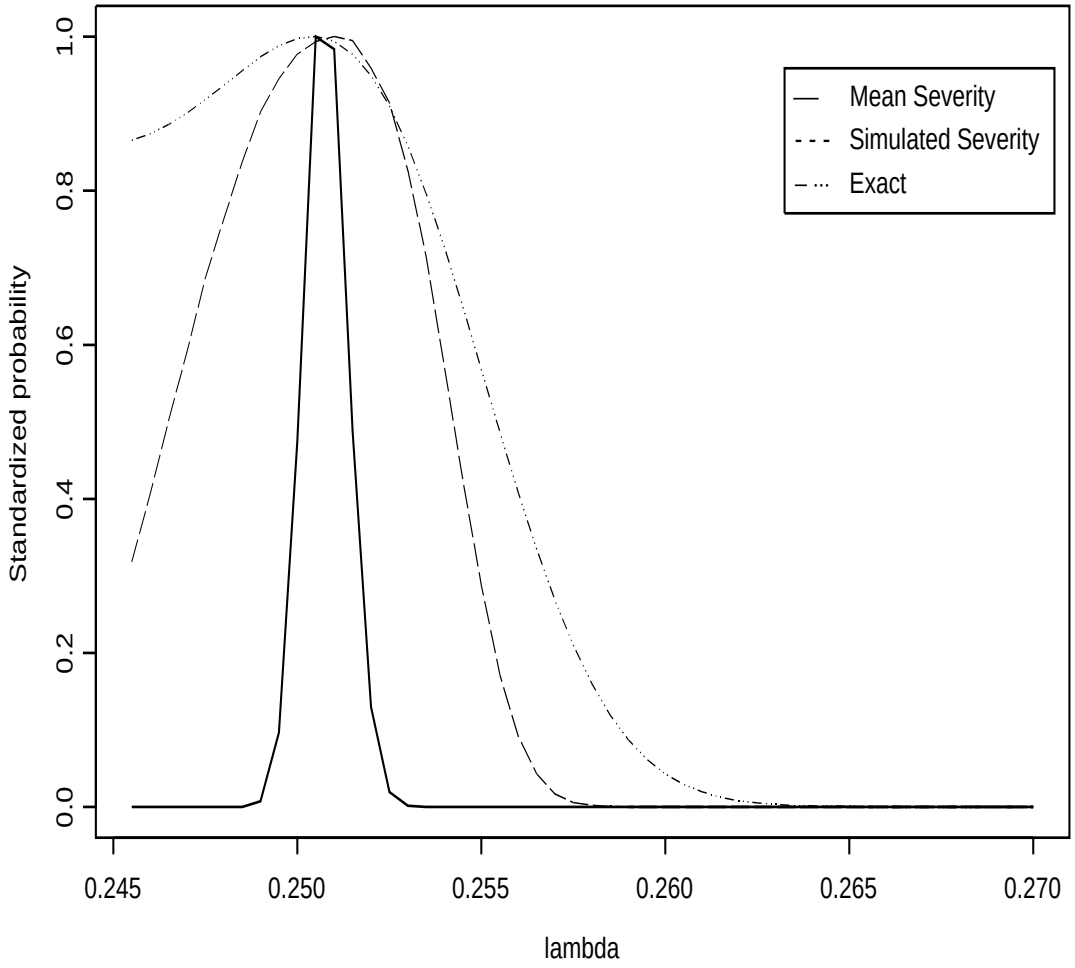
Can also compare exact final size probabilities with those obtained from

$$P(R_\infty = x) = \binom{N}{x} p^x (1-p)^{N-x},$$

$$p = 1 - \exp(-\lambda_G T_\infty / N).$$

Can (i) simulate T_∞ values; or (ii) more crudely set $T_\infty/N = \mu$, its (approximate) mean.

Likelihood comparison for exact, simulated, and mean severity values.



Comments

Approximation works best away from $R_* = 1$

Could refine by allowing T_∞ to take small values corresponding to minor epidemics

Other auxilliary variable methods

The Future

Model choice methodology

Inference for complex models and datasets