Scalable Inference for Epidemic Models With Individual Level Data

Panayiota Touloupou\textsuperscript{1}, Simon Spencer\textsuperscript{2} and Bärbel Finkenstädt\textsuperscript{2}

\textsuperscript{1}School of Mathematics, University of Birmingham
\textsuperscript{2}Department of Statistics, University of Warwick
Introduction

Bayesian inference for epidemic models

Simulation studies

Epidemics with genetic typing data

Discussion
Introduction
Statistical epidemic modelling

- Insights into dynamics of infectious diseases
  - Prevention.
  - Control spread of the disease.

- Epidemiological data present several challenges
  - Missing data (typically high dimensional).
  - Diagnostic tests imperfect.

- **Statistical inference** for epidemic models is hard
  - Intractable likelihood - need to know missing times.
  - Usual solution: large scale data augmentation MCMC.

- What are the observed data?
- **Household data**: Individuals form groups (e.g. households).
Individual level data

- **Household data**: Individuals form groups (e.g. households).
- **Longitudinal** observations.

![Diagram showing the progression of infection status from Day 1 to Day T with non-infected (black) and infected (red) individuals.](image-url)
• **Household data**: Individuals form groups (e.g. households).

• **Longitudinal** observations.

![Diagram showing individual levels over time with symbols for non-infected and infected days.](diagram)
Individual level data

- **Household data**: Individuals form groups (e.g. households).
- **Longitudinal** observations.
Challenges!

- **GOAL**: Draw inference for the parameters given the model.

- Inference for disease outbreak data is **hard**
  - Missing data $X$ typically very high dimensional.

- Intractable likelihood:
  \[
  \pi(Y | \theta) = \sum_X \pi(Y | X, \theta).
  \]

- Solution:
  - Include the hidden infection status of individuals as a model parameter.
  - Use **MCMC data augmentation**.
Diagram of the Markov discrete time epidemic model. Circles are hidden states and rectangles are observed data. Arrows represent dependencies.
Bayesian inference for epidemic models
Bayesian data augmentation

Initialise: Draw $\theta^{(0)} \sim \pi(\theta)$ and generate $X^{(0)} \sim \pi \left( X \mid \theta^{(0)} \right)$;

for $j = 1, 2, \ldots, J$ do
    Update $\theta^{(j)}$ according to $\pi \left( \theta \mid Y, X^{(j-1)} \right)$;
    Update $X^{(j)}$ according to $\pi \left( X \mid Y, \theta^{(j)} \right)$;
end
Bayesian data augmentation

**MCMC Scheme**

Initialise: Draw $\theta^{(0)} \sim \pi(\theta)$ and generate $X^{(0)} \sim \pi(X | \theta^{(0)})$;

for $j = 1, 2, \ldots, J$ do

   Update $\theta^{(j)}$ according to $\pi(\theta | Y, X^{(j-1)})$;

   Update $X^{(j)}$ according to $\pi(X | Y, \theta^{(j)})$;

end
Existing methods

- **Block Update Method**\(^a\):  
  - Choose one **block of states** for each individual and propose one of 3 possible changes: **Add** or **Remove** a block of infection/clearance or **Move** an endpoint of such a block.

- **Single-Site Method**\(^b\):  
  - Update each **single node** from its full conditional distribution.

- **Forward Filtering Backward Sampling (FFBS)**\(^c\):  
  - Update the **whole hidden process** from its full conditional.
  - Computationally intensive.

---

\(^a\) S. E. F. Spencer et al. “‘Super’ or just ‘above average’? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle”. In: *Journal of The Royal Society Interface* 12 (2015).


Existing methods

- **Block Update Method:**
  - Choose one block of states for each individual and propose one of 3 possible changes: Add or Remove a block of infection/clearance or Move an endpoint of such a block.

- **Single-Site Method:**
  - Update each single node from its full conditional distribution.

- **Forward Filtering Backward Sampling (FFBS):**
  - Update the whole hidden process from its full conditional.
  - Computationally intensive.

**Problem**

Algorithms do not scale well to large populations.
Vanilla FFBS

Reformulate graph:

- $x_t^{[1:C]} = (x_t^{[1]}, x_t^{[2]}, \ldots, x_t^{[C]})$
- $\in X^C = \{1, 2, \ldots, N\}^C$.
- $|x_t^{[1:C]}| = N^C$.
- Update the whole hidden process $X$ from its full conditional:
  \[ X \sim \pi(X | Y, \theta) ; \]
- Computational complexity: $\mathcal{O}(TN^{2C})$.

$N = \text{number of infection states.}$
$C = \text{number of individuals.}$
$T = \text{number of time-points.}$
Proposed method: individual FFBS (iFFBS)

Reformulate graph:

- Modification of FFBS.
- Update one individual at a time by sampling from the full conditional:
  \[ \pi \left( X_{1:T}^c \mid Y, X_{1:T}^{[-c]}, \theta \right). \]
- Computational complexity reduced to \( \mathcal{O}(TCN^3) \).

\[ N = \text{number of infection states.} \]
\[ C = \text{number of individuals.} \]
\[ T = \text{number of time-points.} \]

---

Simulation studies
Application: SIS Markov model

- Stochastic SIS (Susceptible-Infected-Susceptible) transmission model in discrete time.

- \( X_{t}^{[c, p]} \) is the infection state of individual \( c \) in group \( p \) on day \( t \):
  - \( X_{t}^{[c, p]} = 0 \) - susceptible/uninfected.
  - \( X_{t}^{[c, p]} = 1 \) - infected/carrier.

- Susceptible individuals acquire infection via two routes:
  - Direct or indirect transmission from other infected individuals within the group.
  - External transmission; transmission from other environmental sources from outside the group.
The transition probabilities between the states are given by:

\[ 1 - e^{-\alpha - \beta \sum_{i=1}^{N} x_t^{[c,p]}} \]

\[ e^{-\alpha - \beta \sum_{i=1}^{N} x_t^{[c,p]}} \]

where \( \alpha \) and \( \beta \) are the external and within-group transmission rates, respectively, and \( m \) is the mean infection period.

- Individuals are initially infected with probability \( \nu \).
- Tests are assumed to have perfect specificity but imperfect sensitivity.
Comparison of methods: Estimation

Number of infected individuals vs Day for True and Block, True and Single-site, True and fullFFBS, True and iFFBS.
Comparison of methods: Time and ACF

- Time (in seconds)
- ACF per iteration

**Graphs:**
- Individuals in group
- Lag

**Legend:**
- Green: Block
- Red: Single-site
- Cyan: fullFFBS
- Blue: iFFBS
Comparison of methods: Larger population

<table>
<thead>
<tr>
<th>Individuals in group</th>
<th>Method</th>
<th>Relative speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Block</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Single-site</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iFFBS</td>
<td></td>
</tr>
</tbody>
</table>
Epidemics with genetic typing data
Motivating example

- 160 cattle randomly assigned in 20 pens, 8 cattle per pen.

- Two test results for *E. coli* O157:H7:
  - Faecal sample,
  - Recto-Anal Mucosal Swab (RAMS).

- Individuals were sampled 27 times over a 99 day period.

- 12 isolates were randomly selected from each pen to be typed using PFGE.
Multi-strain data

Pen 1 (south)

Animal index

Pen 2 (south)

Pen 6 (north)

Pen 8 (north)

Time (days)

Test  

Faecal (bottom row)  

RAMS (top row)
48 different types (arbitrarily label according to the order in which they appeared in the PFGE typing).

- 24 appeared only once.
- 7 major types (at least 10 RAMS and/or faecal samples).
Multi-Strain epidemic model

- Stochastic multi-state model in discrete time.
- \( X_{t}^{[c,p]} \) unobserved carriage status for animal \( c \) in pen \( p \) on day \( t \).
  - \( X_{t}^{[c,p]} = 0 \): non-carrier.
  - \( X_{t}^{[c,p]} = s, \ s = 1, 2, \ldots, 7 \): carriage of one of the common genotypes.
  - \( X_{t}^{[c,p]} = 8 \): carriage of the remaining genotypes (pooled group).

- Imperfect test sensitivity:
  - Falsely recorded as non-carrier.
  - Misclassified as another genotype.

---

Transitions between the states

- Acquisition rate: \( \lambda_p^s(t) = \alpha_s + \beta_s \sum_{i=1}^{C} 1_{\{X_t^{c,p} = s\}} \)

- Clearance rate: \( \mu_s \)

- Relative colonisation rate in a carrier versus non-carrier: \( \delta \)

Example of an epidemic model with 3 competing types.
Comparing parameters between genetic types

<table>
<thead>
<tr>
<th>Genotype (s)</th>
<th>$\nu_s \times 100$</th>
<th>$\alpha_s \times 100$</th>
<th>$\beta_s \times 100$</th>
<th>$\mu_s \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D (1)</td>
<td>2.909</td>
<td>0.123</td>
<td>0.989</td>
<td>16.104</td>
</tr>
<tr>
<td></td>
<td>(0.314, 5.814)</td>
<td>(0.050, 0.204)</td>
<td>(0.367, 1.693)</td>
<td>(9.713, 23.002)</td>
</tr>
<tr>
<td>J (2)</td>
<td>0.556</td>
<td>0.080</td>
<td>1.222</td>
<td>17.164</td>
</tr>
<tr>
<td></td>
<td>(0.000, 2.455)</td>
<td>(0.024, 0.152)</td>
<td>(0.290, 2.411)</td>
<td>(8.574, 27.164)</td>
</tr>
<tr>
<td>X (3)</td>
<td>0.686</td>
<td>0.122</td>
<td>1.093</td>
<td>13.310</td>
</tr>
<tr>
<td></td>
<td>(0.000, 2.484)</td>
<td>(0.054, 0.203)</td>
<td>(0.473, 1.834)</td>
<td>(8.460, 18.431)</td>
</tr>
<tr>
<td>b (4)</td>
<td>0.261</td>
<td>0.058</td>
<td>0.620</td>
<td>9.789</td>
</tr>
<tr>
<td></td>
<td>(0.000, 1.492)</td>
<td>(0.011, 0.110)</td>
<td>(0.003, 1.259)</td>
<td>(3.268, 17.734)</td>
</tr>
<tr>
<td>d (5)</td>
<td>1.628</td>
<td>0.146</td>
<td>0.693</td>
<td>9.964</td>
</tr>
<tr>
<td></td>
<td>(0.000, 3.896)</td>
<td>(0.063, 0.231)</td>
<td>(0.276, 1.169)</td>
<td>(6.080, 14.202)</td>
</tr>
<tr>
<td>f (6)</td>
<td>0.314</td>
<td>0.059</td>
<td><strong>0.347</strong></td>
<td>6.853</td>
</tr>
<tr>
<td></td>
<td>(0.000, 1.667)</td>
<td>(0.013, 0.118)</td>
<td><strong>(0.000, 0.845)</strong></td>
<td>(0.743, 16.849)</td>
</tr>
<tr>
<td>l (7)</td>
<td>0.955</td>
<td>0.046</td>
<td><strong>1.571</strong></td>
<td>11.767</td>
</tr>
<tr>
<td></td>
<td>(0.000, 2.601)</td>
<td>(0.009, 0.094)</td>
<td><strong>(0.901, 2.345)</strong></td>
<td>(7.268, 17.091)</td>
</tr>
<tr>
<td>Pooled (8)</td>
<td>2.119</td>
<td>0.192</td>
<td>0.723</td>
<td>9.501</td>
</tr>
<tr>
<td></td>
<td>(0.000, 5.443)</td>
<td>(0.086, 0.314)</td>
<td>(0.274, 1.186)</td>
<td>(6.081, 13.002)</td>
</tr>
</tbody>
</table>
## Comparing parameters between genetic types

<table>
<thead>
<tr>
<th>Genotype (s)</th>
<th>$\nu_s \times 100$</th>
<th>$\alpha_s \times 100$</th>
<th>$\beta_s \times 100$</th>
<th>$\mu_s \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D (1)</td>
<td>2.909 (0.314, 5.814)</td>
<td>0.123 (0.050, 0.204)</td>
<td>0.989 (0.367, 1.693)</td>
<td>16.104 (9.713, 23.002)</td>
</tr>
<tr>
<td>J (2)</td>
<td>0.556 (0.000, 2.455)</td>
<td>0.080 (0.024, 0.152)</td>
<td>1.222 (0.290, 2.411)</td>
<td>17.164 (8.574, 27.164)</td>
</tr>
<tr>
<td>X (3)</td>
<td>0.686 (0.000, 2.484)</td>
<td>0.122 (0.054, 0.203)</td>
<td>1.093 (0.473, 1.834)</td>
<td>13.310 (8.460, 18.431)</td>
</tr>
<tr>
<td>b (4)</td>
<td>0.261 (0.000, 1.492)</td>
<td>0.058 (0.011, 0.110)</td>
<td>0.620 (0.003, 1.259)</td>
<td>9.789 (3.268, 17.734)</td>
</tr>
<tr>
<td>d (5)</td>
<td>1.628 (0.000, 3.896)</td>
<td>0.146 (0.063, 0.231)</td>
<td>0.693 (0.276, 1.169)</td>
<td>9.964 (6.080, 14.202)</td>
</tr>
<tr>
<td>f (6)</td>
<td>0.314 (0.000, 1.667)</td>
<td>0.059 (0.013, 0.118)</td>
<td>0.347 (0.000, 0.845)</td>
<td>6.853 (0.743, 16.849)</td>
</tr>
<tr>
<td>l (7)</td>
<td>0.955 (0.000, 2.601)</td>
<td>0.046 (0.009, 0.094)</td>
<td>1.571 (0.901, 2.345)</td>
<td>11.767 (7.268, 17.091)</td>
</tr>
<tr>
<td>Pooled (8)</td>
<td>2.119 (0.000, 5.443)</td>
<td>0.192 (0.086, 0.314)</td>
<td>0.723 (0.274, 1.186)</td>
<td>9.501 (6.081, 13.002)</td>
</tr>
</tbody>
</table>
Rest of the parameters

- The median relative colonisation rate in a carrier versus non-carrier individual is 0.842.

- Test sensitivities:
  - RAMS test: 76%,
  - Faecal test: 46%.

- 81.6% of the common genotypes are correctly classified as the right type.

- 1.2% are misclassified as another common type.

- 17.2% are misclassified as type 8.

- 98% of the observed pooled genotypes 8 are correctly classified as 8.
Posterior probability of infection by type
Simulations: Reconstructing the untyped observations

Within-pen colonization rates

False Positive Rate

Within-pen colonization rates
Discussion
Discussion

- iFFBS algorithm exploits the dependence structure in epidemic data to achieve scalable inference.

- Allows much more complex models to be fitted, e.g. with genetic data (epiPOMS\textsuperscript{3} R package).

- Can reconstruct the genetic type of every infection from surprisingly few typed observations.

- Can be used as a Metropolis-Hastings proposal to fit semi-Markov epidemic models.

- Can be used for scalable model selection (Jake Carson and Simon Spencer).

Extension: Investigating transmission between neighbouring pens

Arrows represent potential transmission routes between infected and a given susceptible individual.
Future work

- Improve the computational efficiency of iFFBS even more (e.g., update subset of individuals).

- Extend the multi-genotype model, e.g.:
  - Co-infection: allow for colonisation by all pairwise combinations of single carriage states,
  - Semi-Markov infection period: Negative Binomial distribution.
THANK YOU!!! Any Questions?

Acknowledgement:

- Simon Spencer
- Bärbel Finkenstädt
- Nigel P. French
- Thomas E. Besser
References


Misclassification Matrices

For the case where a positive RAMS sample was not chosen to be genotyped we have that:

\[
E^{R+} = \begin{bmatrix}
0 & + \\
0 & 1 & 0 \\
& 1 - \theta_R & \theta_R \\
& \vdots & \vdots & \vdots \\
n_s & 1 - \theta_R & \theta_R \\
\end{bmatrix}
\]

where \( \theta_R \) is the sensitivity of the RAMS test and is denoted by

\[
\theta_R = P\left(R_t^{[c,p]} = + \mid X_t^{[c,p]} = r\right).
\]
Misclassification matrices

For a positive sample that was genotyped we introduce additional parameters $\theta_C$, $\theta_S$ and $\theta_U$:

$E^{R_s} =$

\[
\begin{bmatrix}
0 & 1 & \cdots & \cdots & \cdots & \cdots & n_s - 1 & n_s (\text{Type U}) \\
0 & 1 - \theta_R & \theta_C \theta_R & \frac{\theta_S \theta_R}{n_s - 2} & \cdots & \cdots & \frac{\theta_S \theta_R}{n_s - 2} & (1 - \theta_C - \theta_S) \theta_R \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
n_s - 1 & \frac{\theta_S \theta_R}{n_s - 2} & \frac{\theta_S \theta_R}{n_s - 2} & \cdots & \cdots & \theta_C \theta_R & \frac{\theta_S \theta_R}{n_s - 2} & (1 - \theta_C - \theta_S) \theta_R \\
n_s (\text{Type U}) & 1 - \theta_R & \theta_U \theta_R & \frac{\theta_U \theta_R}{n_s - 1} & \cdots & \cdots & \frac{\theta_U \theta_R}{n_s - 1} & (1 - \theta_U) \theta_R
\end{bmatrix}
\]

such that, for all $r \neq 0$, the probabilities

$e_{r,0}^{R_s} = \Pr(R_t^{[c, p]} = 0 \mid X_t^{[c, p]} = r) = 1 - \theta_R$ and $\sum_{s=1}^{n_s} e_{r,s}^{R_s} = \theta_R$. 