Bayesian inference and model selection for stochastic epidemics and other coupled hidden Markov models

(with special attention to epidemics of *Escherichia coli* O157:H7 in cattle)

Simon Spencer

3rd May 2016

WARWICK
THE UNIVERSITY OF WARWICK

Warwick Statistics

Acknowledgements

Panayiota Touloupou

Bärbel Finkenstädt Rand
Peter Neal
TJ McKinley
Nigel French, Tom Besser and
Rowland Cobbold

Outline

1. Introduction

2. Bayesian inference for epidemics

3. Model selection for epidemics

4. Scalable inference for epidemics

5. Conclusion

Introduction

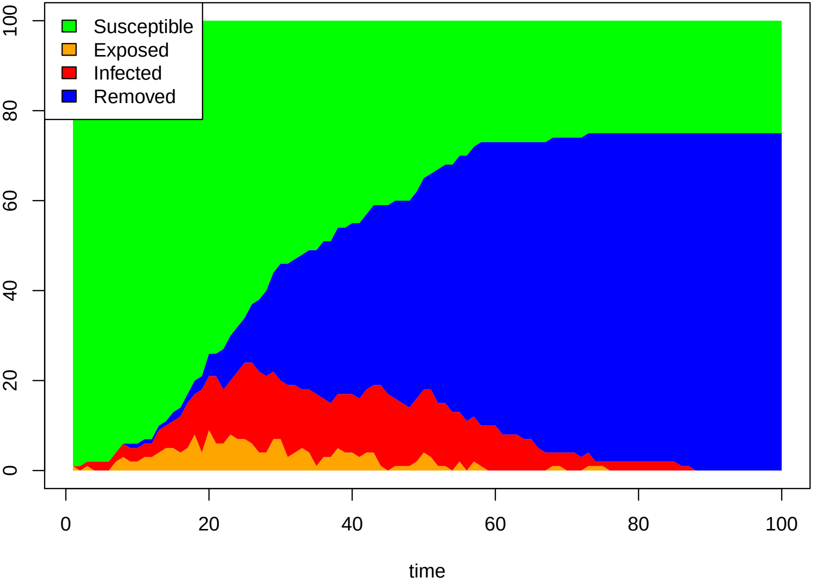
Introduction

A typical epidemic model:

Susceptible → *Exposed* → *Infected* → *Removed*

Infections occur according to an inhomogeneous Poisson process with rate $\propto S(t)I(t)$.

A simulation



Comments

- Statistical inference for epidemic models is hard.
- Intractable likelihood – need to know infection times.
- Usual solution: large scale data augmentation MCMC.
- What are the observed data?

Epidemic data

- Historically: final size (single number).
    - Final size in many sub-populations, e.g. households.
- Markov models: removal times.
    - Who is removed is not needed / recorded.

Epidemic data

- Historically: final size (single number).
    - Final size in many sub-populations, e.g. households.
- Markov models: removal times.
    - Who is removed is not needed / recorded.
- Individual level diagnostic test results.
    - To be realistic, tests are imperfect.
    - Temporal resolution of 1 day.

Epidemic data

- Historically: final size (single number).
  - Final size in many sub-populations, e.g. households.

- Markov models: removal times.
  - Who is removed is not needed / recorded.

- Individual level diagnostic test results.
  - To be realistic, tests are imperfect.
  - Temporal resolution of 1 day.

⇒ View epidemic as **hidden Markov model**

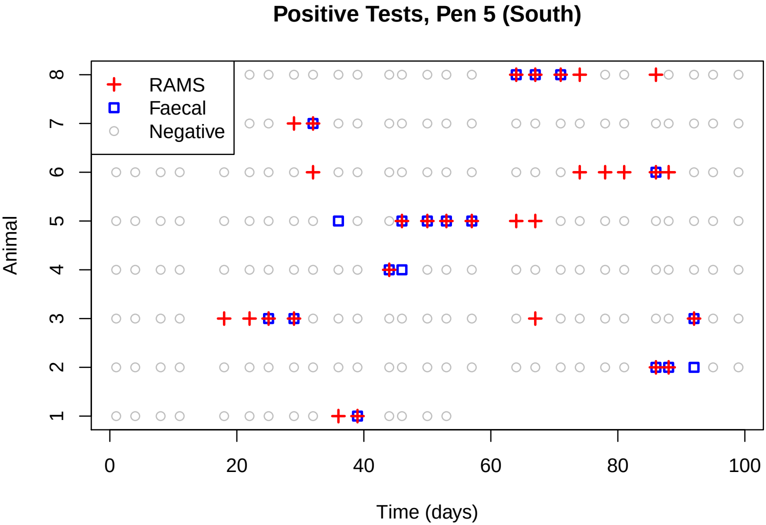
Motivating example: *Escherichia coli* O157

- *E. coli* O157 is a highly pathogenic form of *Escherichia coli*.
- It can cause severe gastroentestinal illness, haemorrhagic diarrhoea and even death.
- Outbreaks and endemic cases are associated with food, water or direct contact with infected animals.
- Cattle are the main reservoir.
- Additional economic burden due to impacts on trade.

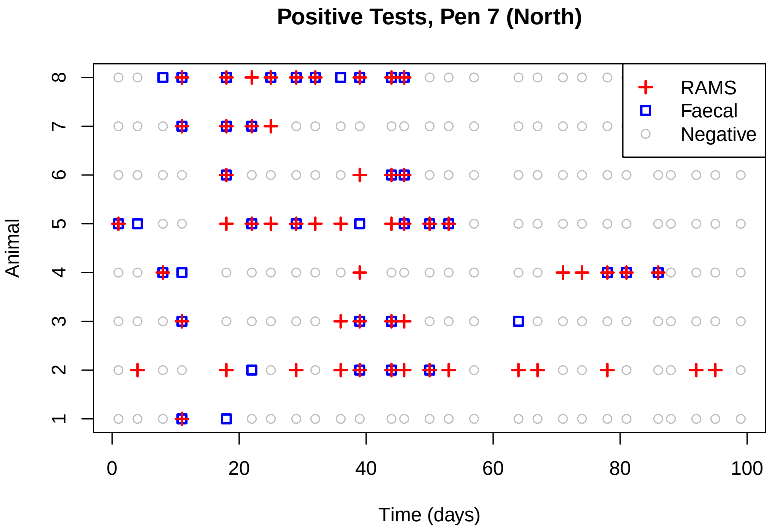
Study design

- Natural colonization and faecal excretion of *E. coli* O157 in commercial feedlot.
- 20 pens containing 8 calves were sampled 27 times over a 99 day period.
- Each sampling event included a faecal pat sample and a recto-anal mucosal swab (RAMS).
- Tests were assumed to have perfect *specificity* but imperfect *sensitivity*.

Patterns of infection



Patterns of infection



Bayesian inference for epidemics

Bayesian inference for epidemics

- Intractable likelihood: $\pi(\boldsymbol{y}|\boldsymbol{\theta})$.
- Need to impute infection status of individuals $\boldsymbol{x}$ for augmented likelihood $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.
- Missing data $\boldsymbol{x}$ typically very high dimensional.

Updating the infection status

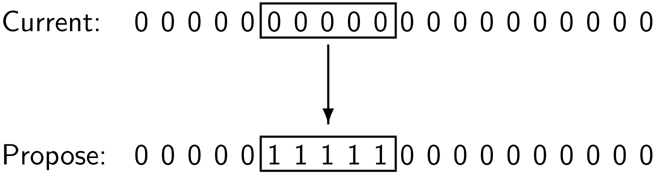
- Standard method by O'Neill and Roberts (1999) involves 3 steps:

  1. **Add** a period of infection
  2. **Remove** a period of infection
  3. **Move** an end-point of a period of infection

- This method was designed for SIR models (where individuals can't be infected twice).

- Easily adapted to discrete time models.

Add a period of infection

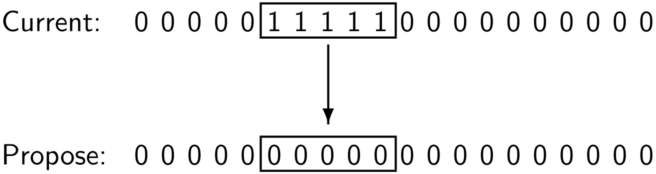


Current: 0 0 0 0 0 $\boxed{0\ 0\ 0\ 0\ 0}$ 0 0 0 0 0 0 0 0 0 0

Propose: 0 0 0 0 0 $\boxed{1\ 1\ 1\ 1\ 1}$ 0 0 0 0 0 0 0 0 0 0

1. Choose a block of zeros at random.
2. Propose changing zeros to ones.
3. Accept or reject based on ratio of posteriors.

Remove a period of infection

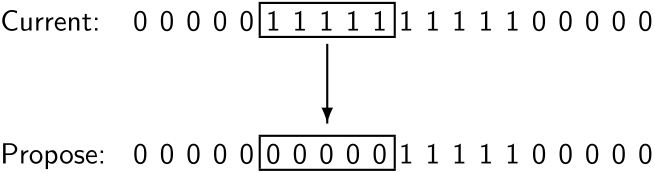


Current: 0 0 0 0 0 $\boxed{1\ 1\ 1\ 1\ 1}$ 0 0 0 0 0 0 0 0 0 0

Propose: 0 0 0 0 0 $\boxed{0\ 0\ 0\ 0\ 0}$ 0 0 0 0 0 0 0 0 0 0

1. Choose a complete block of ones.
2. Propose changing ones to zeros.
3. Accept or reject based on ratio of posteriors.

Move an endpoint



Current: $0\ 0\ 0\ 0\ 0\ \boxed{1\ 1\ 1\ 1\ 1}\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0$

Propose: $0\ 0\ 0\ 0\ 0\ \boxed{0\ 0\ 0\ 0\ 0}\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0$

1. Choose an endpoint of a block of ones.
2. Propose a new location for that endpoint.
3. Accept or reject based on ratio of posteriors.

Some pros and cons

- ✓ Considerably fast
- ✓ Can handle non-Markov models
- ✗ Most of the hidden states are not updated
- ✗ High degree of autocorrelation
  - Slow mixing of the chain and long run length
- ✗ Tuning of the maximum block length required.

Alternative approach: FFBS

- Discrete time epidemic is a hidden Markov model.
- Gibbs step: sample from the full condition distribution of the hidden states.

- Use Forward Filtering Backward Sampling algorithm (Carter and Kohn, 1994).

Some pros and cons

✓ Very good mixing of the MCMC chains

✓ No tuning required

✗ Computationally intensive

- At each timepoint we need to calculate N^C summations
- $\mathcal{O}(TN^{2C})$

✗ High memory requirements

- All T forward variables must be stored
- The transition matrix is of dimension $N^C \times N^C$

N = number of infection states (e.g. 2)
C = number of cows (e.g. 8)
T = number of timepoints (e.g. 99)

Example: SIS model

- Stochastic SIS (Susceptible-Infected-Susceptible) transmission model in discrete time.[1]

- $X_{p,i,t}$ infection status for animal i in pen p on day t.
    - $X_{p,i,t} = 1$ – infected/colonized.
    - $X_{p,i,t} = 0$ – uninfected/susceptible.

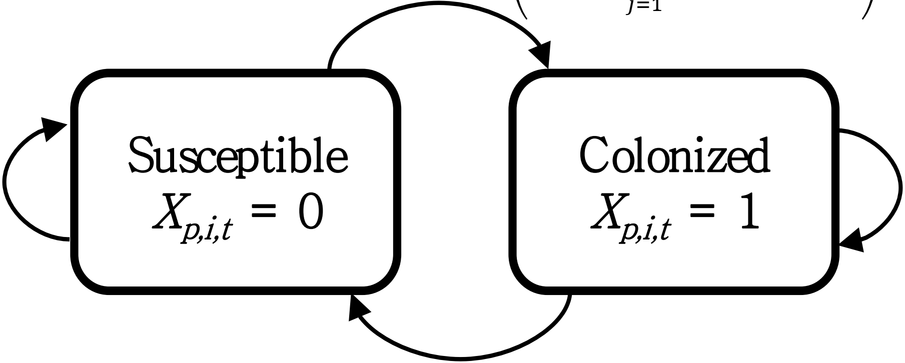
- We treat $X_{p,i,t}$ as missing data and infer it using MCMC.

- Epidemic model parameters updated via Metropolis-Hastings and test sensitivities updated using Gibbs.

[1] Spencer *et al.* (2015) 'Super' or just 'above average'? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle. *Interface* **12**, 20150446.

Colonization probability:

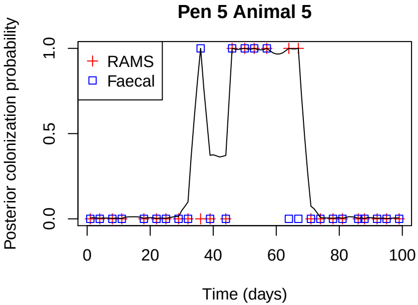
$$\mathrm{P}\left(X_{p,i,t+1} = 1 \middle| X_{p,i,t} = 0\right) = 1 - \exp\left(-\alpha - \beta \sum_{j=1}^{8} X_{p,j,t}\, \rho^{\mathbb{I}(S_{p,j,t} > \tau)}\right)$$



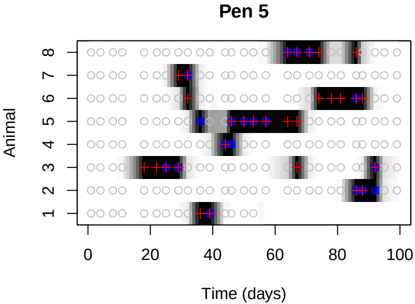
Colonization duration: $\mathrm{NegativeBinomial}(r, \mu)$

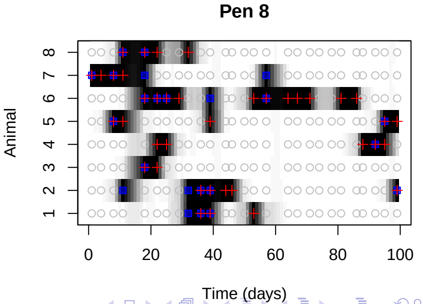
Pens: $p = 1 \cdots 20$ Animals: $i = 1 \cdots 8$ Time: $t = 1 \cdots 99$ days

Example: Posterior infection probabilities



- We can calculate the posterior infection probability for every day of the study.





Model selection for epidemics

Model selection for epidemics

A lot of epidemiologically interesting questions take the form of model selection questions.

- What is the transmission mechanism of this disease?
- Do infected individuals really exhibit an exposed period?
- Do water troughs spread *E. coli* O157?

Posterior probabilities and marginal likelihoods

Would like the posterior probability in favour of model i.

$$\mathrm{P}(M_i|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|M_i)\mathrm{P}(M_i)}{\sum_j \pi(\boldsymbol{y}|M_j)\mathrm{P}(M_j)}$$

Posterior probabilities and marginal likelihoods

Would like the posterior probability in favour of model i.

$$\mathrm{P}(M_i|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|M_i)\mathrm{P}(M_i)}{\sum_j \pi(\boldsymbol{y}|M_j)\mathrm{P}(M_j)}$$

Equivalently, the Bayes factor comparing models i and j.

$$B_{ij} = \frac{\pi(\boldsymbol{y}|M_i)}{\pi(\boldsymbol{y}|M_j)}$$

Posterior probabilities and marginal likelihoods

Would like the posterior probability in favour of model i.

$$\mathrm{P}(M_i|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|M_i)\mathrm{P}(M_i)}{\sum_j \pi(\boldsymbol{y}|M_j)\mathrm{P}(M_j)}$$

Equivalently, the Bayes factor comparing models i and j.

$$B_{ij} = \frac{\pi(\boldsymbol{y}|M_i)}{\pi(\boldsymbol{y}|M_j)}$$

All we need is the marginal likelihood,

$$\pi(\boldsymbol{y}|M_i) = \int \pi(\boldsymbol{y}|\boldsymbol{\theta}, M_i)\pi(\boldsymbol{\theta}|M_i)\,\mathrm{d}\boldsymbol{\theta}$$

but how can we calculate it?

Marginal likelihood estimation

- Many existing approaches:
  - Chib's method
  - Power posteriors
  - Harmonic mean
  - Bridge sampling

- Most direct approach: importance sampling.

- Use asymptotic normality of the posterior to find efficient proposal.

- But how to deal with the missing data?

Dr Peter Neal

Marginal likelihood estimation using importance sampling

1. Run MCMC as usual.
2. Fit normal distribution to posterior samples[2] $\Rightarrow q(\boldsymbol{\theta})$.
3. Draw N samples from $q(\boldsymbol{\theta})$.

$$\pi(\boldsymbol{y}) = \int \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}.$$

[2] To avoid problems, make q overdispersed relative to the posterior.

Marginal likelihood estimation using importance sampling

1. Run MCMC as usual.
2. Fit normal distribution to posterior samples[2] $\Rightarrow q(\boldsymbol{\theta})$.
3. Draw N samples from $q(\boldsymbol{\theta})$.

$$\pi(\boldsymbol{y}) \approx \sum_{i=1}^{N} \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}.$$

[2] To avoid problems, make q overdispersed relative to the posterior.

Marginal likelihood estimation with missing data

1. Run MCMC as usual.
2. Fit normal distribution to posterior samples → $q(\boldsymbol{\theta})$.
3. Draw N samples from $q(\boldsymbol{\theta})$.
4. For each sampled $\boldsymbol{\theta}_i$ draw missing data $\boldsymbol{x}_i$ from the full conditional using FFBS.

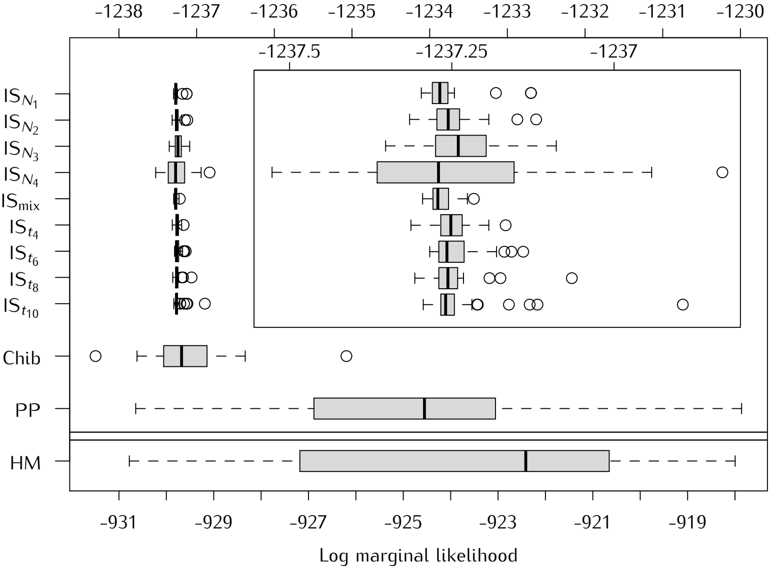
$$\pi(\boldsymbol{y}) \approx \sum_{i=1}^{N} \frac{\pi(\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\theta}_i)\ \pi(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\ \pi(\boldsymbol{\theta}_i)}{\pi(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{\theta}_i)\ q(\boldsymbol{\theta}_i)}.$$

Simulation study: pneumococcol carriage

- Panayiota performed a thorough simulation study[3] based on Melegaro *at al.* (2004).
- Household based longitudinal study on carriage of *Streptococcus Pneumoniae*.
- Data consist of repeated diagnostic tests.
- Multi-type model with 11 parameters, 2600 observed data and 6500 missing data.

[3] Touloupou *et al.* (2016) Model comparison with missing data using MCMC and importance sampling. arXiv 1512.04743

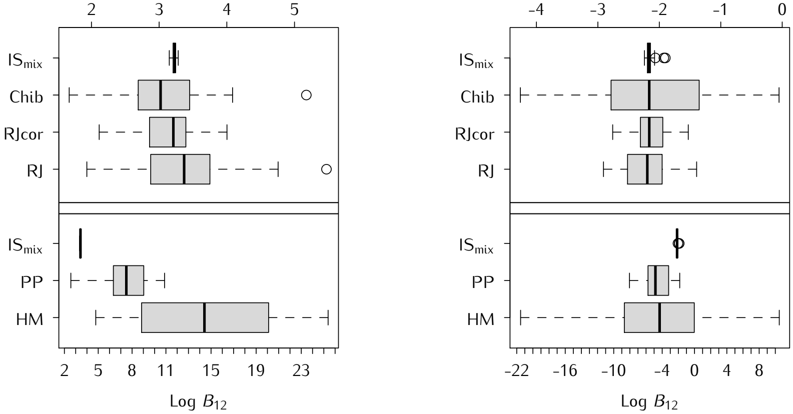
Results: marginal likelihood estimation



Results: Bayes factor estimation

Do adults and children acquire infection at the same rate?

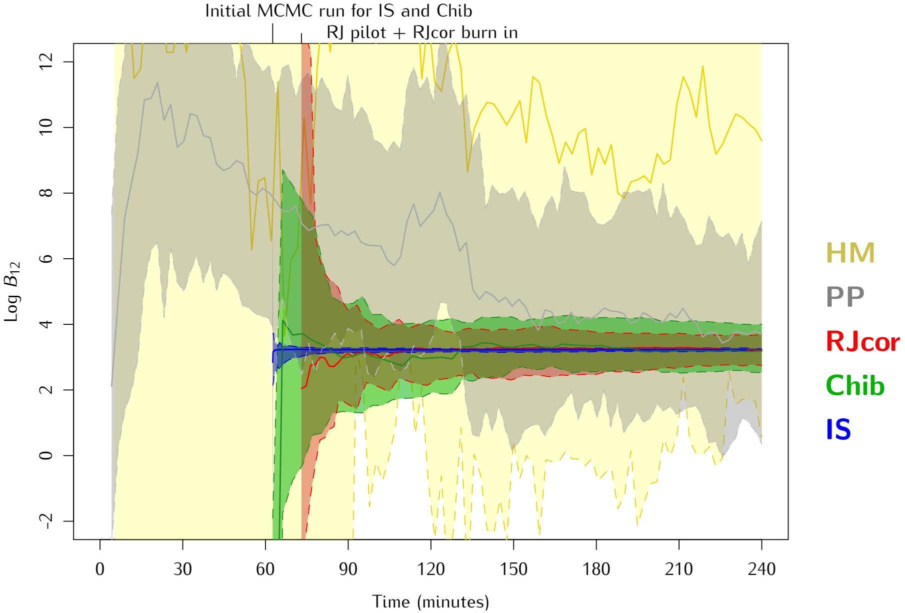
- $M_1 : k_A \neq k_C$
- $M_2 : k_A = k_C$



(a) Data simulated from model M_1

(b) Data simulated from model M_2

Results: Evolution of the log Bayes factor

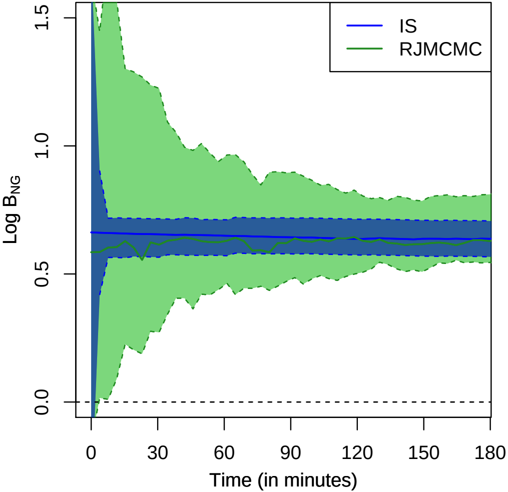


Application 1: *E. coli* O157 in feedlot cattle

Do animals develop immunity over time?

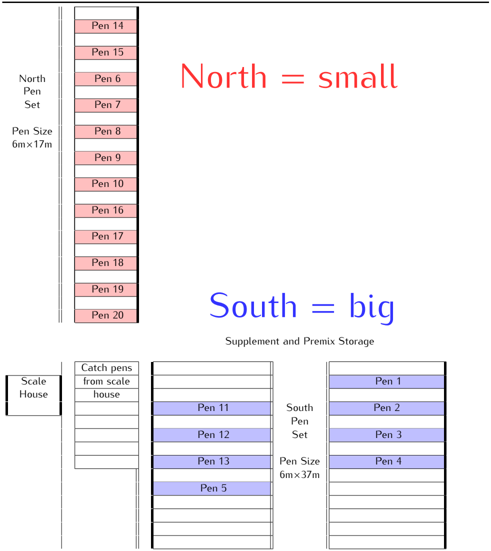
- We compare two models for infection period:
  - Geometric: lack of memory.
  - Negative Binomial: probability of recovery depends on duration of infection.
- The Negative Binomial is a generalisation of the Geometric:
  - Setting Negative Binomial dispersion parameter $\kappa = 1$ leads to Geometric.

Application 1: Results



- **RJMCMC** and **IS** agree on the estimate of the Bayes factor
- **IS** estimator: faster convergence
- Bayes factor supports the Negative Binomial model
- The longer the colonization, the greater the probability of clearance – may indicate an immune response in the host

Application 2: Role of pen area/location



North = small

South = big

Application 2: Role of pen area/location

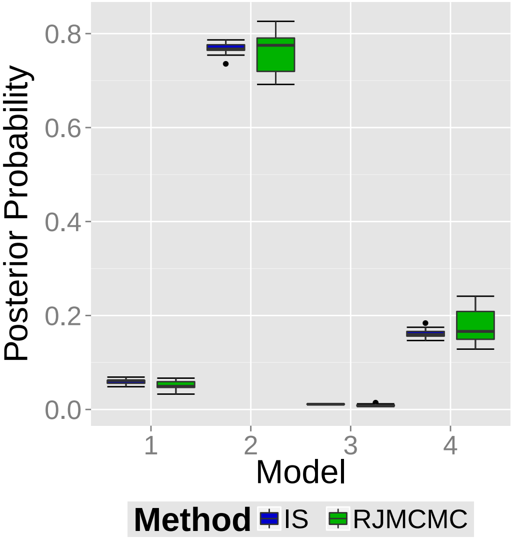
Do north and south pens have different risk of infection?

- Allow different external (α_s, α_n) and/or within-pen (β_s, β_n) transmission rates.

- Candidate models:

	External		Within-pen	
Model	North	South	North	South
1	α_n	α_s	β_n	β_s
2	α	α	β_n	β_s
3	α_n	α_s	β	β
4	α	α	β	β

Application 2: Posterior probabilities

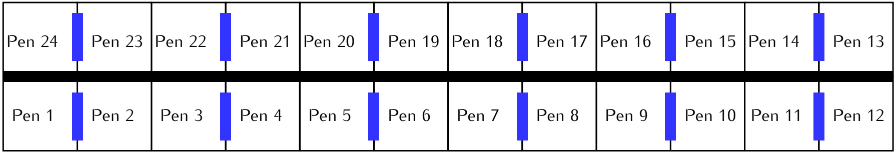


- **RJMCMC** and **IS** provide identical conclusions.
- Evidence to support different within-pen transmission rates.
- Animals in smaller pens more at risk of within-pen infection

Application 3: Investigating transmission between pens

Additional dataset: pens adjacent in a 12 × 2 rectangular grid.

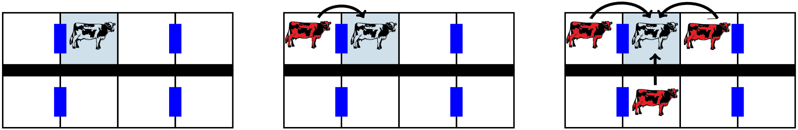
- No direct contact across **feed buck**.
- Shared **waterers** between pairs of adjacent pens.



Pen 24	Pen 23	Pen 22	Pen 21	Pen 20	Pen 19	Pen 18	Pen 17	Pen 16	Pen 15	Pen 14	Pen 13
Pen 1	Pen 2	Pen 3	Pen 4	Pen 5	Pen 6	Pen 7	Pen 8	Pen 9	Pen 10	Pen 11	Pen 12

Application 3: Investigating transmission between pens

Do waterers spread infection?

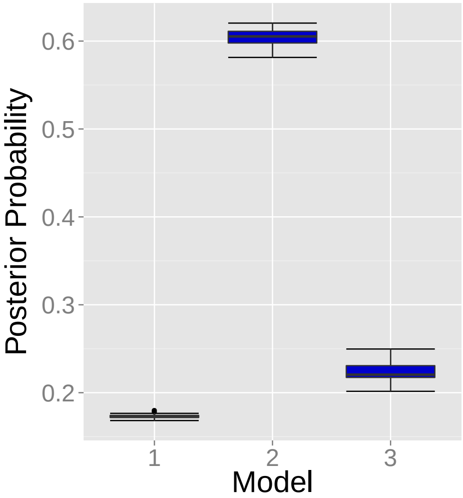


(a) Model 1: No contacts between pens

(b) Model 2: Transmission via a waterer

(c) Model 3: Transmission via any boundary

Application 3: Posterior probabilities



- **RJMCMC**: hard to design jump mechanism
- Using **IS** results still possible.
- Evidence for transmission between pens sharing a waterer rather than another boundary.

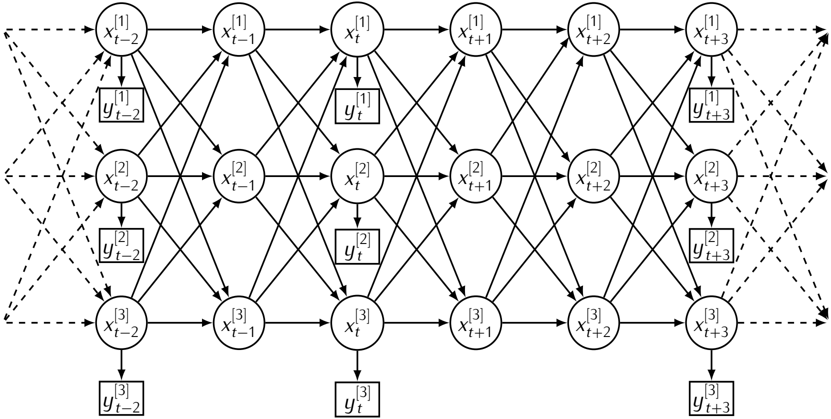
Scalable inference for epidemics

Scalable inference for epidemics

- Thus far we have been doing inference for small populations.
  - Households
  - Pens
- The FFBS algorithm scales very badly with population size.
- We would like an inference method that scales better with population size.

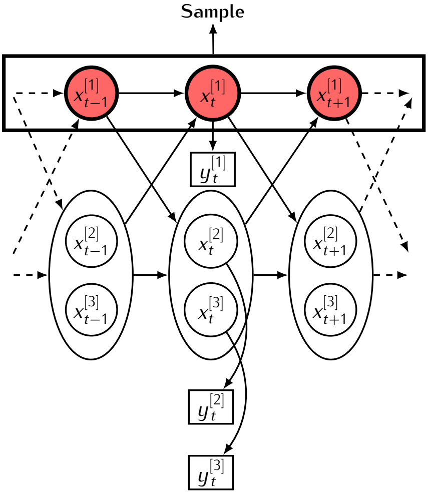
Graphical representation

Diagram of the Markovian epidemic model. Circles are hidden states and rectangles are observed data. Arrows represent dependencies.



A new approach – the iFFBS algorithm

Reformulate graph:



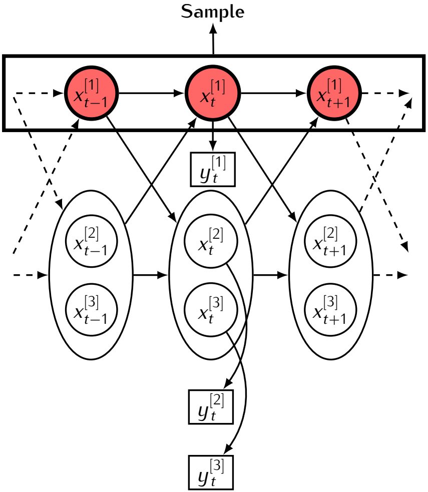
Update one individual at a time by sampling from the full conditional:

$$\mathrm{P}(\boldsymbol{x}_{1:T}^{[c]} \mid \boldsymbol{y}_{1:T}^{[1:C]}, \boldsymbol{x}_{1:T}^{[-c]}, \boldsymbol{\theta}).$$

⇒ View as **coupled** hidden Markov model

A new approach – the iFFBS algorithm

Reformulate graph:



Update one individual at a time by sampling from the full conditional:

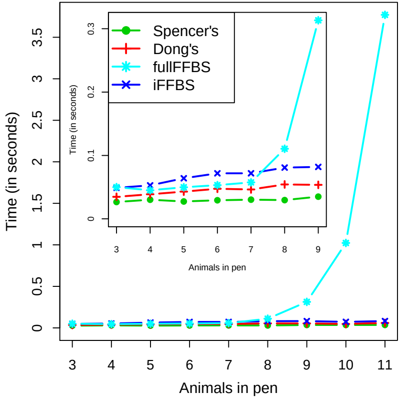
$$\mathrm{P}(\boldsymbol{x}_{1:T}^{[c]} \mid \boldsymbol{y}_{1:T}^{[1:C]}, \boldsymbol{x}_{1:T}^{[-c]}, \boldsymbol{\theta}).$$

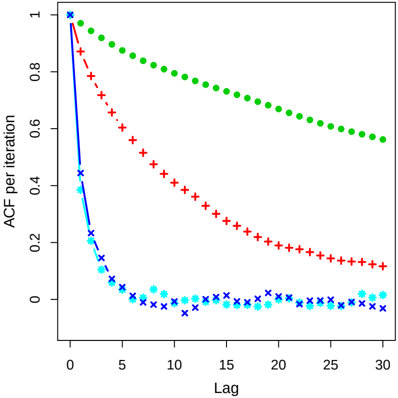
⇒ View as **coupled** hidden Markov model

- Computational complexity reduced from $\mathcal{O}(TN^{2C})$ to $\mathcal{O}(TCN^2)$.

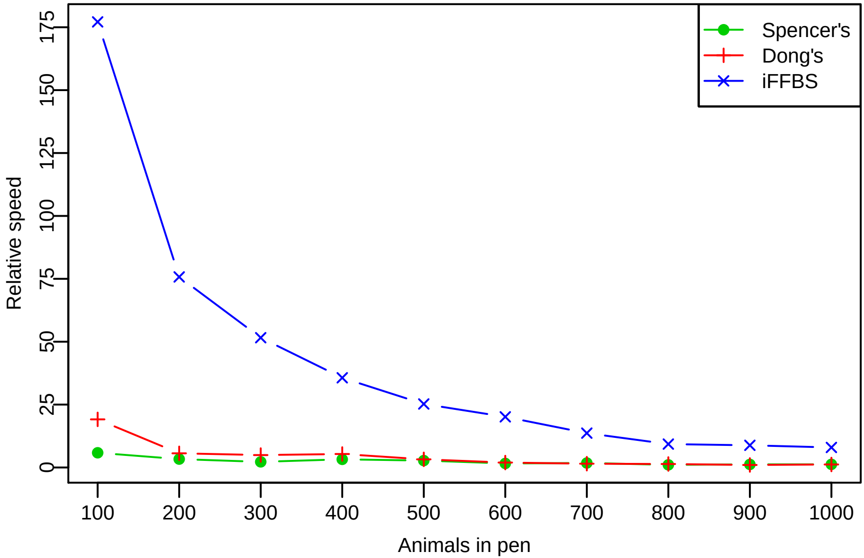
N = number of infection states (e.g. 2)
C = number of cows (e.g. 8)
T = number of timepoints (e.g. 99)

Comparison of methods





Larger populations



Conclusion

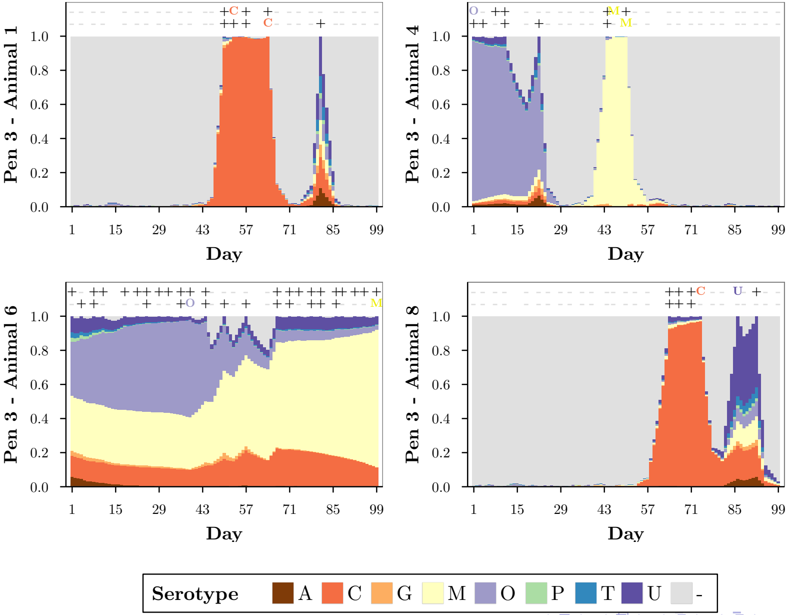
Conclusion

- FFBS algorithm generates better mixing MCMC for parameter inference.
- Unlocks direct approach to marginal likelihood estimation.
- Allows important epidemiological questions to be answered via model selection.
- iFFBS can perform inference in large populations – exploits dependence structure in epidemic data.

What I didn't say

- All of this work (and much more!) has been done by Panayiota.
- FFBS and iFFBS can also be used as a Metropolis-Hastings proposal to fit non-Markovian epidemic models.
- Can we do model selection with iFFBS?
- Power of iFFBS allows more complex models to be fitted, e.g. multi-strain epidemic models.

Current work


Pen 3 - Animal 1
Pen 3 - Animal 4
Pen 3 - Animal 6
Pen 3 - Animal 8
Day
C
C
O
M
M
O
M
C
U
Serotype
A
C
G
M
O
P
T
U