Bayesian Model Selection For Partially Observed Epidemic Models

Panayiota Touloupou

joint work with Simon E.F. Spencer, Bärbel Finkenstädt Rand, Peter Neal, Trevelyan J. McKinley

CRiSM Workshop: Estimating Constants

April 21, 2016

Outline

1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

Outline

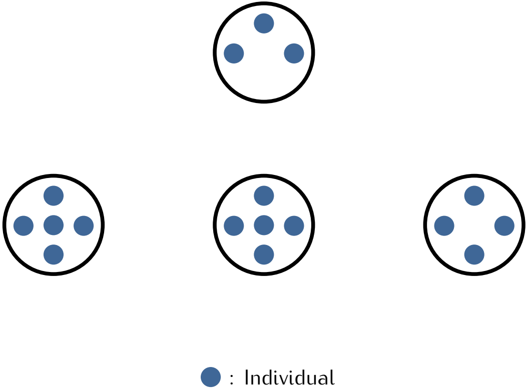
1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

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
- Model selection
  - Each model an epidemiologically important hypothesis

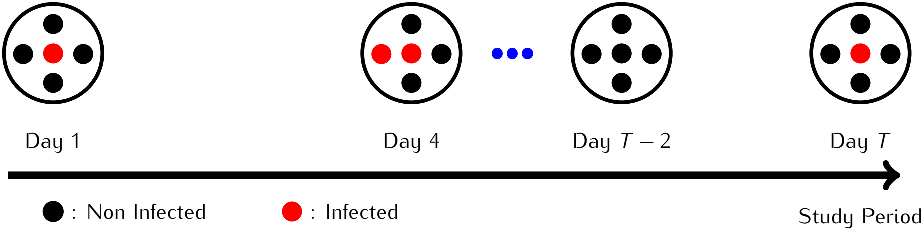
Our Setup

- Longitudinal observations
- Individuals form groups (e.g. households)



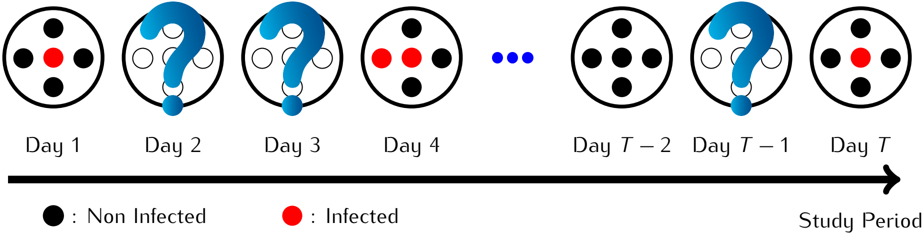
Our Setup

- Longitudinal observations
- Individuals form groups (e.g. households)



Our Setup

- Longitudinal observations
- Individuals form groups (e.g. households)



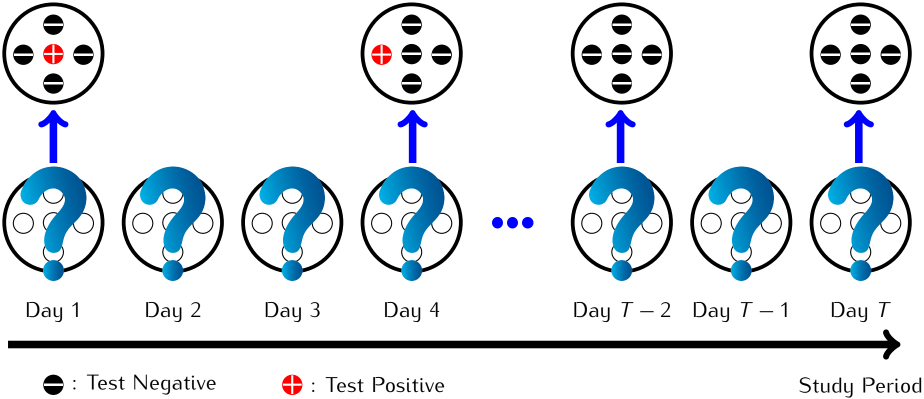
Day 1 Day 2 Day 3 Day 4 Day $T-2$ Day $T-1$ Day T

● : Non Infected 🔴 : Infected

Study Period

Our Setup

- Longitudinal observations
- Individuals form groups (e.g. households)



Objective

- Analysis of this type of data can be challenging
  - ➢ Times of acquiring and clearing infection are unobserved
  - ➢ Intractable likelihood - need to know missing times
  - ➢ Usual solution: large scale data augmentation MCMC

- **Bayesian model selection**
  - ➢ Evidence in favour of candidate models
  - ➢ Each model an epidemiologically important hypothesis

Objectives:

- Develop statistical tools for comparison of competing hypotheses
- Special attention on missing data

Outline

1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

Model Selection For Epidemics

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

- What is the transmission mechanism of the disease?
- Do individuals develop immunity over time?
- 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

- Most direct approach: **Importance sampling**
  - Use asymptotic normality of the posterior to find efficient proposal
- Many existing other approaches:
  - Harmonic mean
  - Chib's methods
  - Power posteriors
  - Bridge sampling

Importance Sampling[1]

1. Obtain samples from the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ with MCMC
2. Use MCMC samples to inform the proposal distribution $\Rightarrow q(\boldsymbol{\theta})$
3. Draw N samples from $q(\boldsymbol{\theta})$
4. Estimate the marginal likelihood by

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

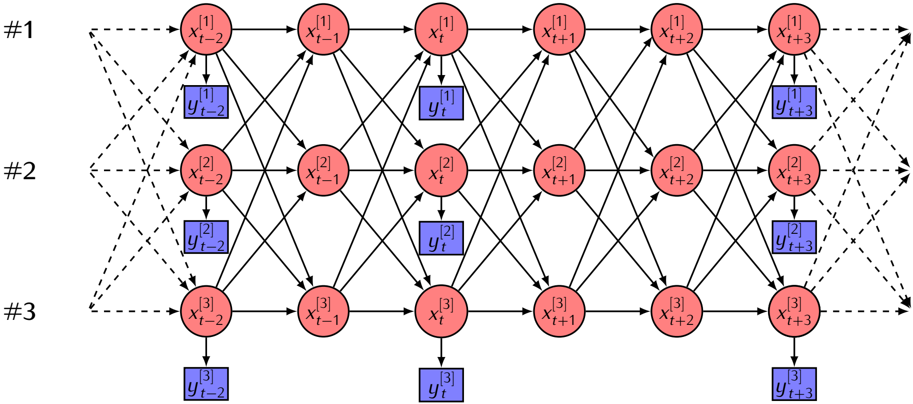
[1] Clyde et al. (2007). Current Challenges in Bayesian Model Choice

Missing Data!

But how to deal with the missing data?

Missing Data!

But how to deal with the missing data?



Importance Sampling With Missing Data

1. Obtain samples from the joint posterior $\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y})$ with MCMC

2. Use MCMC samples to inform the proposal distribution $\Rightarrow q(\boldsymbol{\theta})$

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 Forward Filtering Backward Sampling

5. Estimate the marginal likelihood by

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

Harmonic Mean[2]

- The marginal likelihood $\pi(\boldsymbol{y})$ can be approximated

$$\widehat{\pi}_{HM}(\boldsymbol{y}) = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi(\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\theta}_i)}\right]^{-1}$$

  based on N draws $(\boldsymbol{x}_1, \boldsymbol{\theta}_1), (\boldsymbol{x}_2, \boldsymbol{\theta}_2), \ldots, (\boldsymbol{x}_N, \boldsymbol{\theta}_N)$ from the joint posterior $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$.

- Can be computed directly from MCMC output
- Asymptotically consistent
- Exhibit large or even infinite variance

[2] Newton M.A. and Raftery A.E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap *J. R. Stat. Soc. Ser. B. Stat. Methodol*, **56**, 3–48

Chib's Methods[3]

- Based on the observation that

$$\pi(\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\,\pi(\boldsymbol{x},\boldsymbol{\theta})}{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}$$

for fixed $\boldsymbol{\theta}^*, \boldsymbol{x}^*$ (high-density posterior point) the log marginal likelihood can be estimated by

$$\log \widehat{\pi}_{\text{Chib}}(\boldsymbol{y}) = \log \pi(\boldsymbol{y}|\boldsymbol{x}^*,\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{x}^*,\boldsymbol{\theta}^*) - \log \widehat{\pi}(\boldsymbol{x}^*,\boldsymbol{\theta}^*|\boldsymbol{y})$$

$\Longrightarrow$ is estimated by breaking the parameter vector into appropriate blocks

- Required a separate MCMC run to calculate each block

[3] Chib S. (1995) Marginal likelihood from the Gibbs output *J. Amer. Statist. Assoc*, **90**, 1313–1321. Chib S. and Jeliazkov I. (2001) Marginal likelihood from the MH output *J. Amer. Statist. Assoc*, **96**, 270–281

Power Posteriors[4]

- Power Posterior defined as

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}, t) \propto \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta})^{t}\, \pi(\boldsymbol{x}, \boldsymbol{\theta})$$

where $t \in [0, 1]$ is a temperature parameter

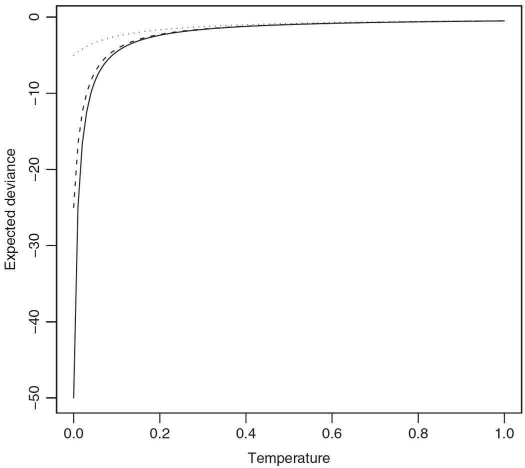
- The log of the marginal likelihood can be represented by

$$\log \pi(\boldsymbol{y}) = \int_0^1 \mathrm{E}_{\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}, t} \Big\{ \log \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \Big\}\, dt$$

$\Longrightarrow$ is calculated numerically by discretising $0 = t_0 < t_1 < \cdots < t_n = 1$, and then using trapezium rule.

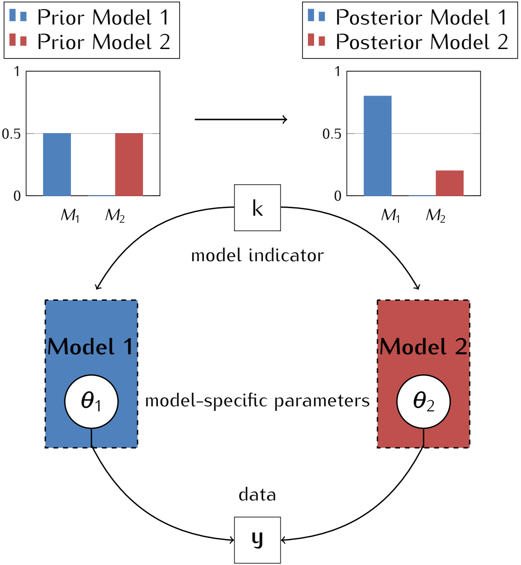
[4] Friel N. and Pettitt A. N. (2008) Marginal likelihood estimation via power posteriors *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70**, 589–607

Power Posteriors: Example



- Obtain samples from the power posterior at each temperature t_i
- Variability depends
  - Number of t_i's
  - Spacing of t_i's
  - Number of MCMC samples
- Large number $\implies$ more computational effort

Reversible Jump MCMC



model indicator

model-specific parameters

data

Outline

1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

Simulation Study: Pnemonococcal Carriage[5]

- Household based longitudinal study on carriage of Streptococcus Pneumoniae

- Diagnostic tests obtained every 4 weeks
  - 10 months period
  - Classified as Negative / Positive

- The population is divided into two age groups:
  - Children 👧👦: under 5 years old
  - Adults 👩👨 : over 5 years old

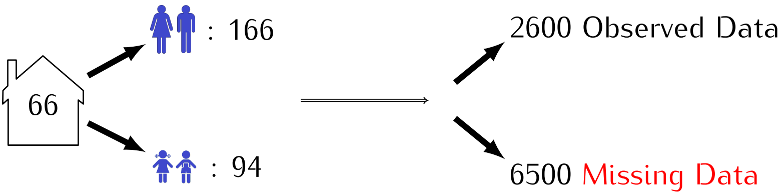
[5] Touloupou et al. (2016) Model comparison with missing data using MCMC and importance sampling. arXiv 1512.04743

Model Details[6]

- Discrete time Susceptible-Infected-Susceptible model
- The transition probabilities age group i dependent:

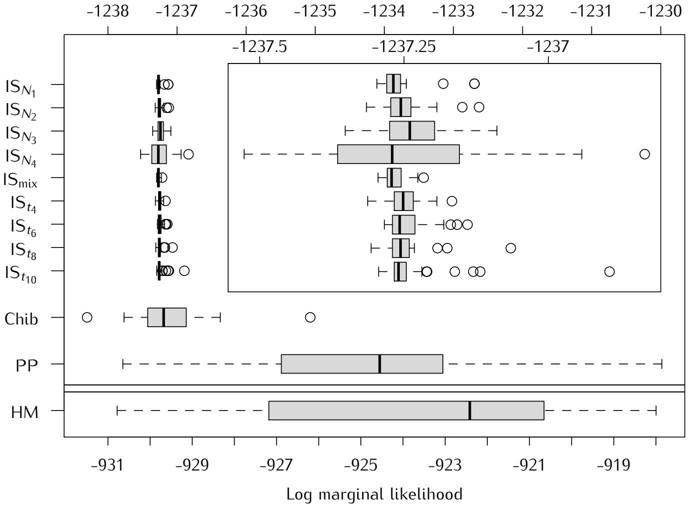
$$P_i(\mathsf{S} \longrightarrow \mathsf{I})_{\delta_t} = 1 - \exp\left\{-\left(k_i + \frac{\beta_{Ci}\,\mathsf{I}_C(t) + \beta_{Ai}\,\mathsf{I}_A(t)}{(z-1)^w}\right)\cdot \delta t\right\}$$

$$P_i(\mathsf{I} \longrightarrow \mathsf{S})_{\delta_t} = 1 - \exp\left(-\mu_i \cdot \delta t\right)$$

- 

[6] Melegaro et al. (2004) Estimating the transmission parameters of pneumococcal carriage in households. *Epidemiology and Infection*, **132**,

Results: Marginal Likelihood Estimation



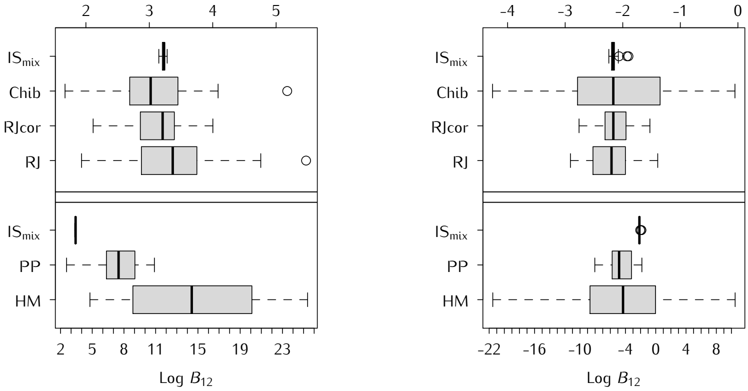
- IS_{N_j} : $N(\boldsymbol{\mu}, j\,\boldsymbol{\Sigma})$ - IS_{t_d} : $t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ - IS_{mix} : $0.95 \times N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.05\pi(\boldsymbol{\theta})$ - $\boldsymbol{\mu}, \boldsymbol{\Sigma}$: from MCMC

Heterogeneity In Community Acquisition Rates

Do adults and children acquire infection at the same rate?

- We compare two models:
  - $\mathcal{M}_1 : k_A \neq k_C$
  - $\mathcal{M}_2 : k_A = k_C$

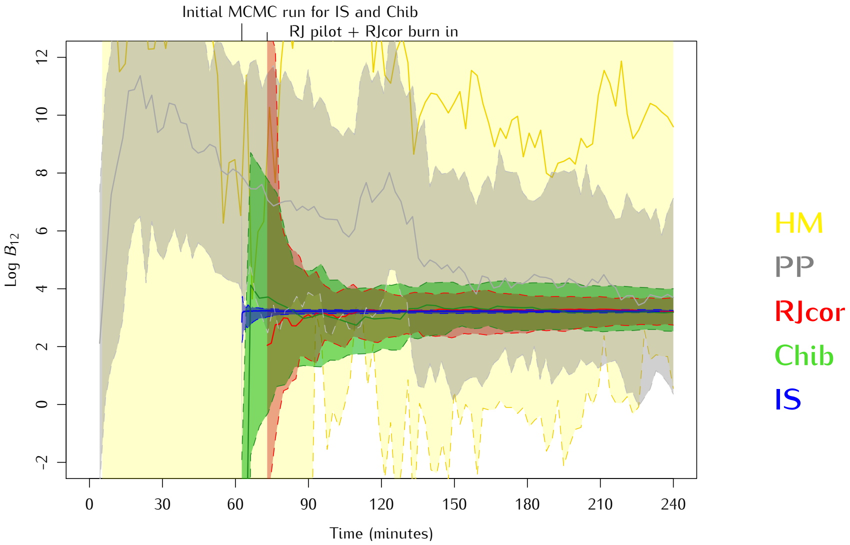
Results: Bayes Factor Estimation



(a) Data simulated from model $\mathcal{M}_1$

(b) Data simulated from model $\mathcal{M}_2$

Results: Evolution Of The Log Bayes Factor



Outline

1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

Study Designs

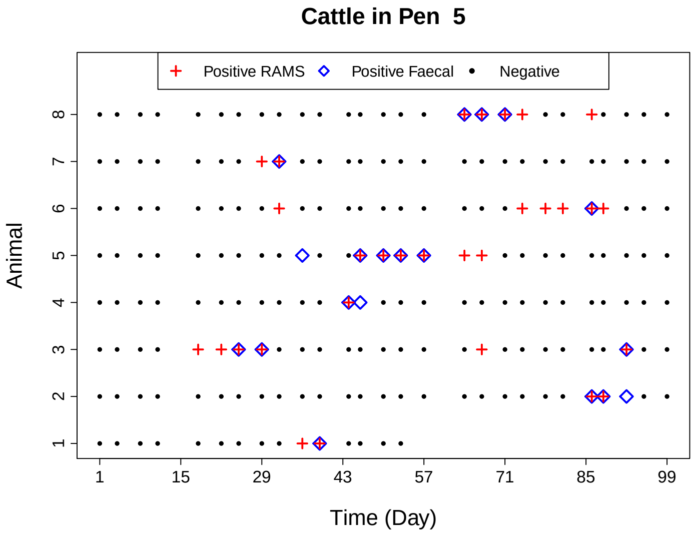
- Two longitudinal studies of *E. coli* O157:H7

	Dataset 1	Dataset 2
Subjects	160 cattle	168 cattle
Study duration	14 weeks	22 weeks
Sampling interval	2 times/week	14 days

- Each sampling event included a
  - Faecal pat sample
  - Recto-anal mucosal swab (RAMS)

- Tests were assumed to have perfect specificity but imperfect sensitivity

Patterns Of Infection



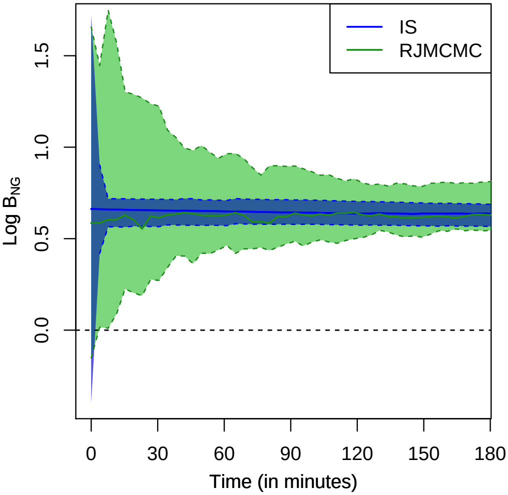
Application 1: E. Coli O157 In Feedlot Cattle

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.

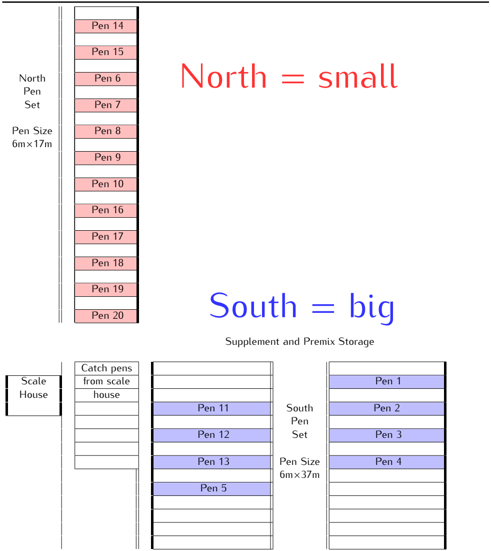
- The Negative Binomial is a generalisation of the Geometric:
    - Setting Negative Binomial dispersion parameter $\kappa = 1$ leads to Geometric.

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

North Pen Set

Pen Size 6m×17m

Pen 14
Pen 15
Pen 6
Pen 7
Pen 8
Pen 9
Pen 10
Pen 16
Pen 17
Pen 18
Pen 19
Pen 20

Supplement and Premix Storage

Scale House

Catch pens from scale house

Pen 11
Pen 12
Pen 13
Pen 5

South Pen Set

Pen Size 6m×37m

Pen 1
Pen 2
Pen 3
Pen 4

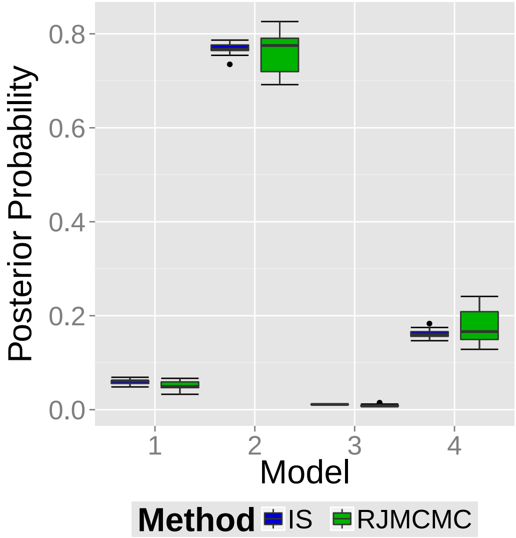
Application 2: Role Of Pen Area/Location

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

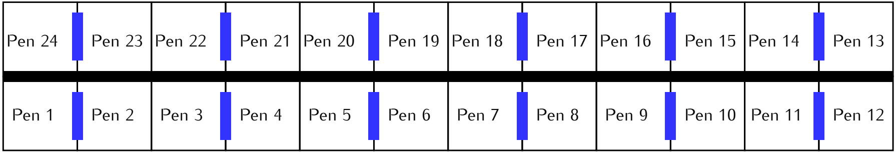


- RJMCMC and IS provide identical conclusions.
- Evidence to support different within-pen transmission rates.
- Animals in smaller pens more at risk of within-pen infection

Application 3: Investigating Transmission Between Pens

Dataset 2: pens adjacent in a 12 × 2 rectangular grid.

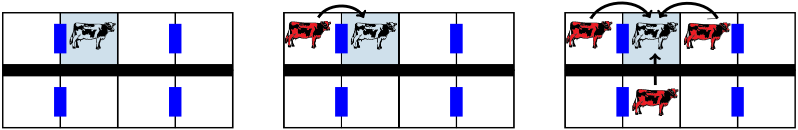
- No direct contact across **feed buck**.
- Shared **waterers** between pairs of adjacent pens.



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?

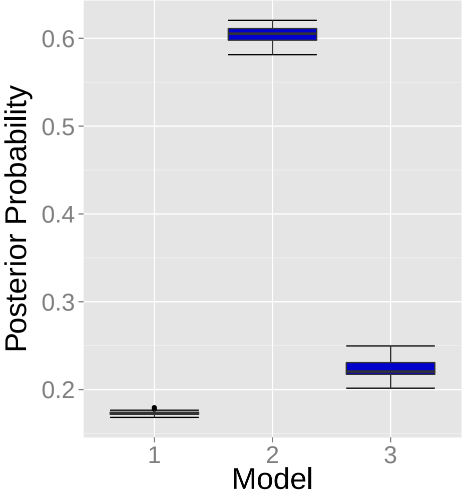


(a) Model 1: No contacts between pens

(b) Model 2: Transmission via a waterer

(c) Model 3: Transmission via any boundary

Application 3: Posterior Probabilities



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

Outline

1. Motivation
2. Methods
3. Simulation Studies
4. Real Data Analysis
5. Conclusions

Concluding Remarks

- Show how IS can be used to test epidemiological questions of interest
- In this study the importance sampling estimator outperformed existing tools
  - ➢ Smallest Monte Carlo error
- Importance sampling approach very easy to implement and trivially parallelisable
- Bayes factors depend on choice of prior
  - ➢ Simulations needed to avoid Lindley's paradox

Variations/Extensions

- When the full conditional is not available we use a related full conditional
  - IS corrects for not using the true full conditional
- My collaborator Peter Neal used the particle filtering to estimate $\pi(\boldsymbol{x}|\boldsymbol{\theta})$
- We recently applied Bridge Sampling for estimating the marginal likelihood
  - IS a special case
  - Slightly reduced variances
  - We use IS due to ease of implementation

Thanks for Listening!

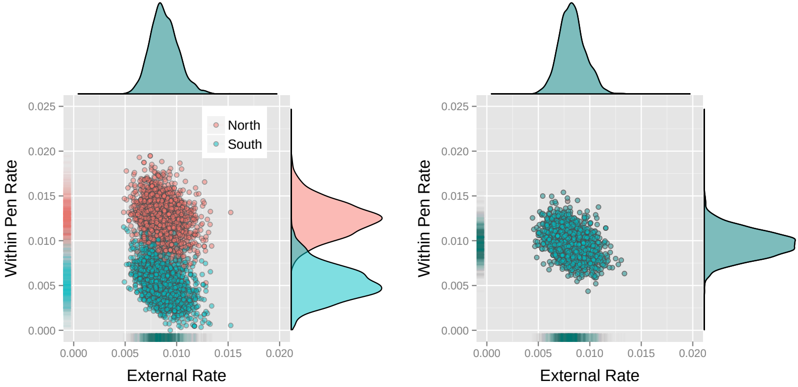
Appendix

Results – Posterior summaries of model parameters

Parameter	Symbol	Geometric	Negative Binomial
External transmission probability	$1 - e^{-\alpha}$	0.0090 [0.0064, 0.0117]	0.0081 [0.0057, 0.0109]
Internal transmission probability	$1 - e^{-\beta}$	0.0107 [0.0077, 0.0141]	0.0102 [0.0073, 0.0137]
Mean period of infection	m	8.9942 [7.7460, 10.4369]	9.9740 [7.1977, 10.6487]
Shape parameter	κ	—	1.6245 [0.8361, 2.8972]
Initial probability of infection	μ	0.1001 [0.0568, 0.1545]	0.0997 [0.0557, 0.1546]
Sensitivity of RAJ test	θ_R	0.7750 [0.7304, 0.8156]	0.7771 [0.7311, 0.8203]
Sensitivity of faecal test	θ_F	0.4639 [0.4206, 0.5073]	0.4657 [0.4213, 0.5097]

- Posterior mean of the parameters of each model along with the 95% credible interval in brackets.

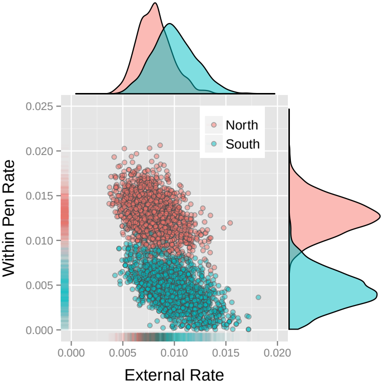
Parameter Estimation



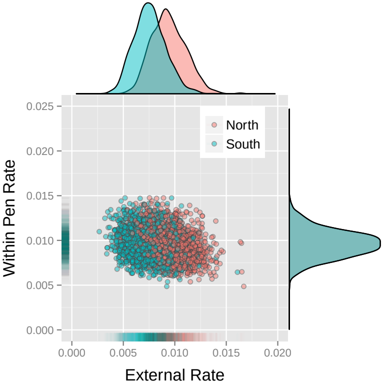
(d) Model 2 - Posterior Prob 0.77

(e) Model 4 - Posterior Prob 0.16

Parameter Estimation



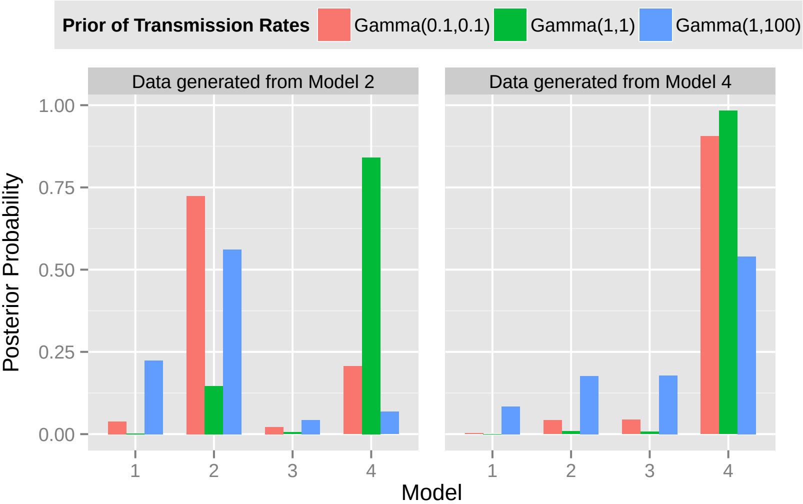
(f) Model 1 - Posterior Prob 0.06



(g) Model 3 - Posterior Prob 0.01

The Choice of Prior Matters!

Simulation study: Heterogeneity in Transmission Rates Among Pens


Prior of Transmission Rates
🟥 Gamma(0.1,0.1)
🟩 Gamma(1,1)
🟦 Gamma(1,100)
Data generated from Model 2
Data generated from Model 4
Posterior Probability
Model