Scalable Inference For Epidemic Models With Individual Level Data

Panayiota Touloupou[1], Simon Spencer[2] and Bärbel Finkenstädt[2]

[1]School of Mathematics, University of Birmingham

[2] Department of Statistics, University of Warwick

Statistics Seminar Series

AUEB

22 April, 2021

UNIVERSITY OF BIRMINGHAM

Overview

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.

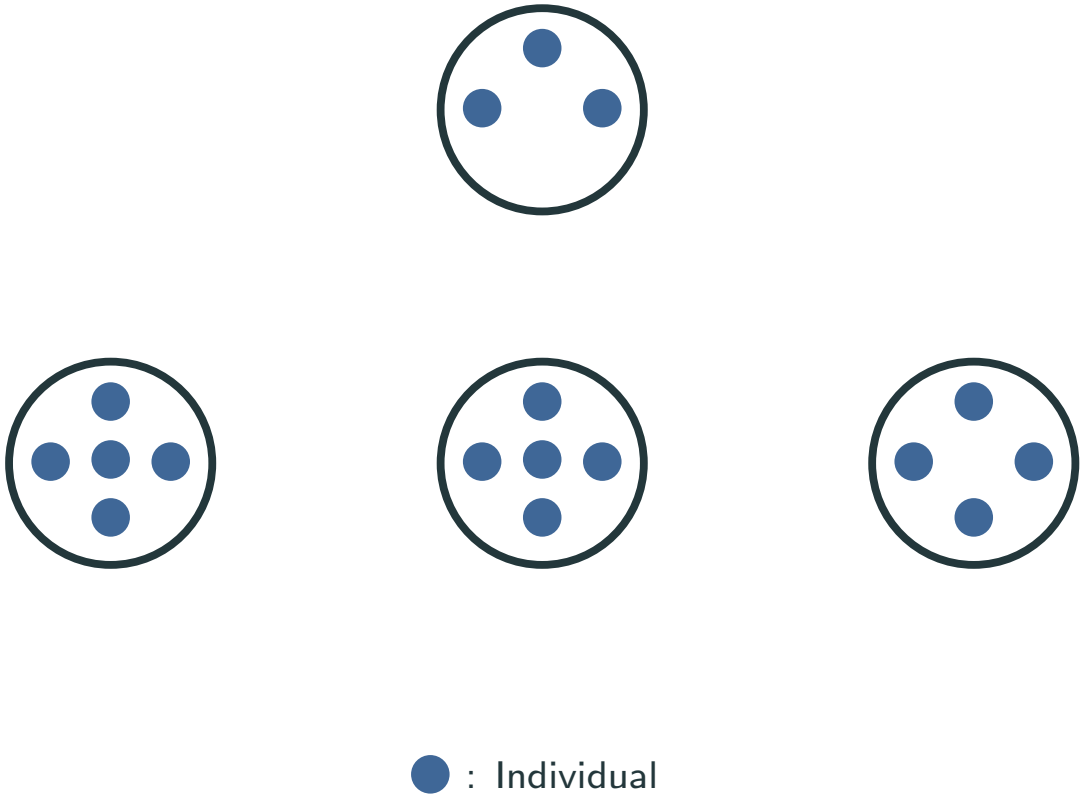
- Epidemiological data present several challenges
  - ➢ Missing data (typically high dimensional).
  - ➢ Diagnostic tests imperfect.

- **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?

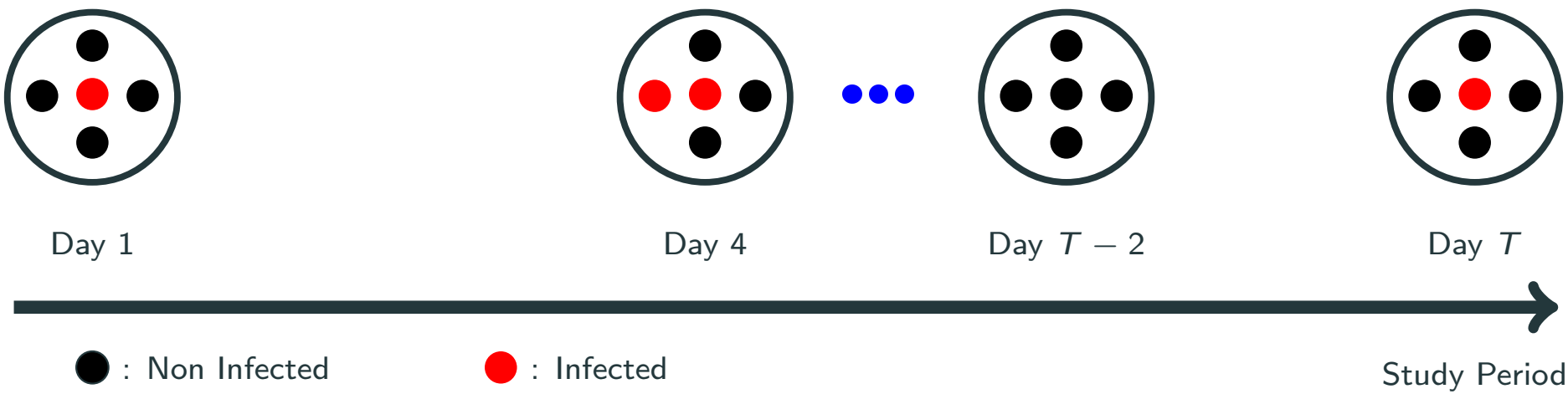
Individual level data

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



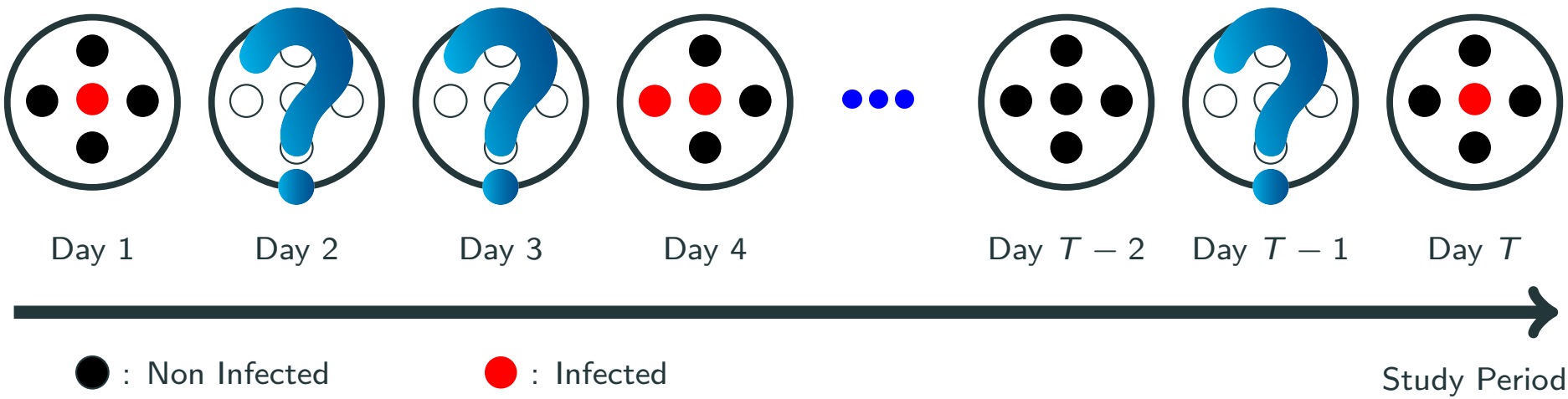
Individual level data

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



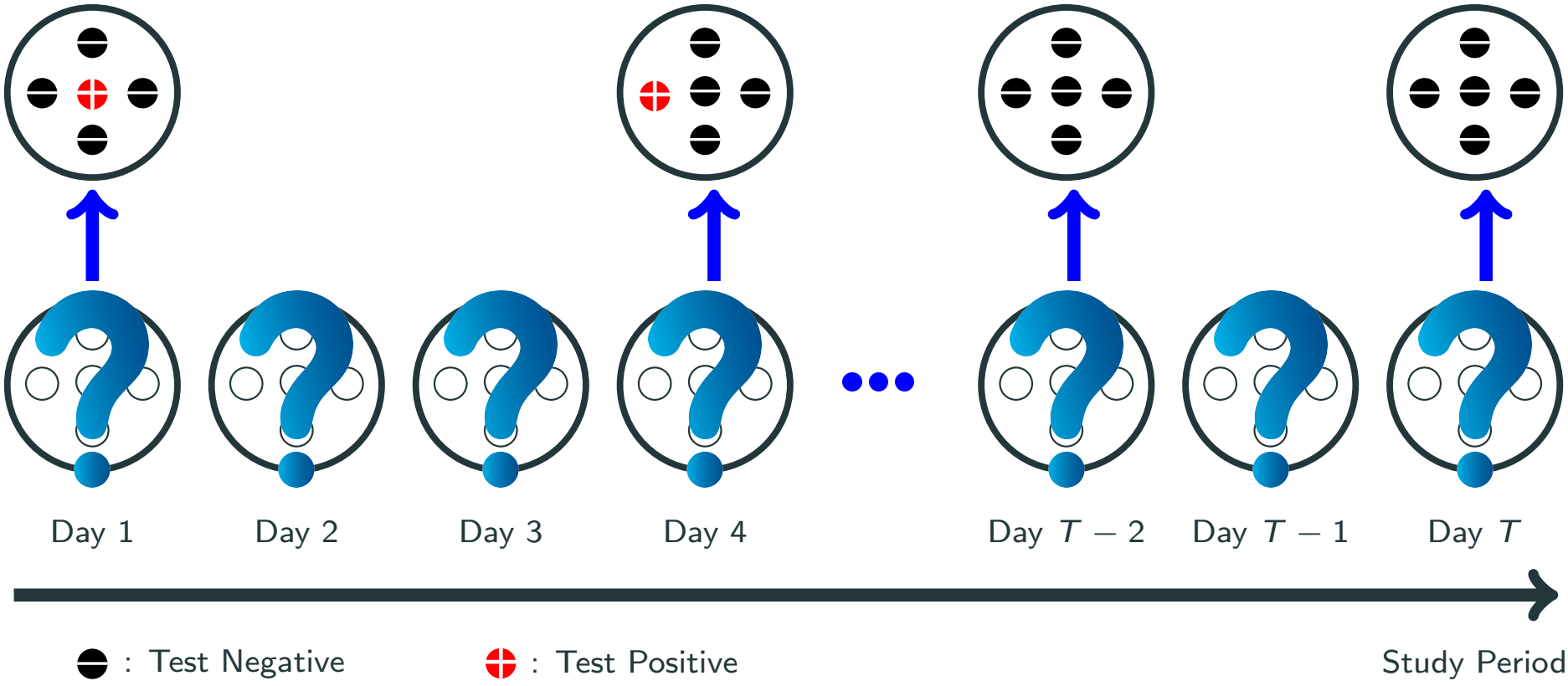
Individual level data

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



Individual level data

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



Challenges!

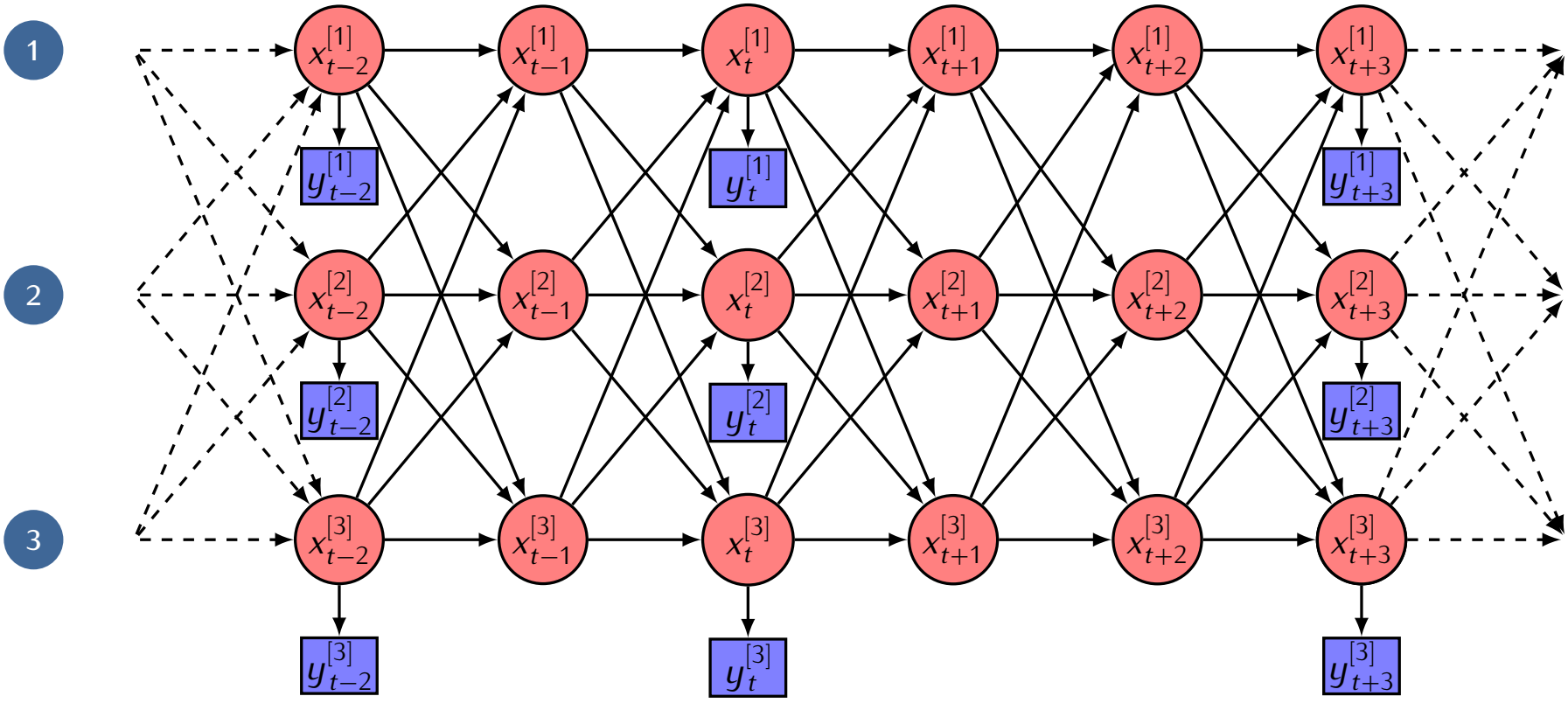
- **GOAL**: Draw inference for the parameters given the model.
- Inference for disease outbreak data is **hard**
  - ➢ **Missing data** $\boldsymbol{X}$ typically very high dimensional.
- Intractable likelihood:

$$\pi(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{X}} \pi(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}).$$

- Solution:
  - ➢ Include the hidden infection status of individuals as a model parameter.
  - ➢ Use **MCMC data augmentation**.

Graphical representation

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

MCMC Scheme

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

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

Update $\boldsymbol{\theta}^{(j)}$ according to $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}^{(j-1)}\right)$;

Update $\boldsymbol{X}^{(j)}$ according to $\pi\left(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(j)}\right)$;

end

Bayesian data augmentation

MCMC Scheme

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

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

Update $\theta^{(j)}$ according to $\pi\left(\theta \mid \boldsymbol{Y}, \boldsymbol{X}^{(j-1)}\right)$;

Update $\boldsymbol{X}^{(j)}$ according to $\pi\left(\boldsymbol{X} \mid \boldsymbol{Y}, \theta^{(j)}\right)$;

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).

[b] W. Dong, A. Pentland, and K. A. Heller. "Graph-coupled HMMs for modeling the spread of infection". In: *arXiv preprint arXiv:1210.4864* (2012).

[c] C. K. Carter and R. Kohn. "On Gibbs Sampling for State Space Models". In: *Biometrika* 3 (1994).

Existing methods

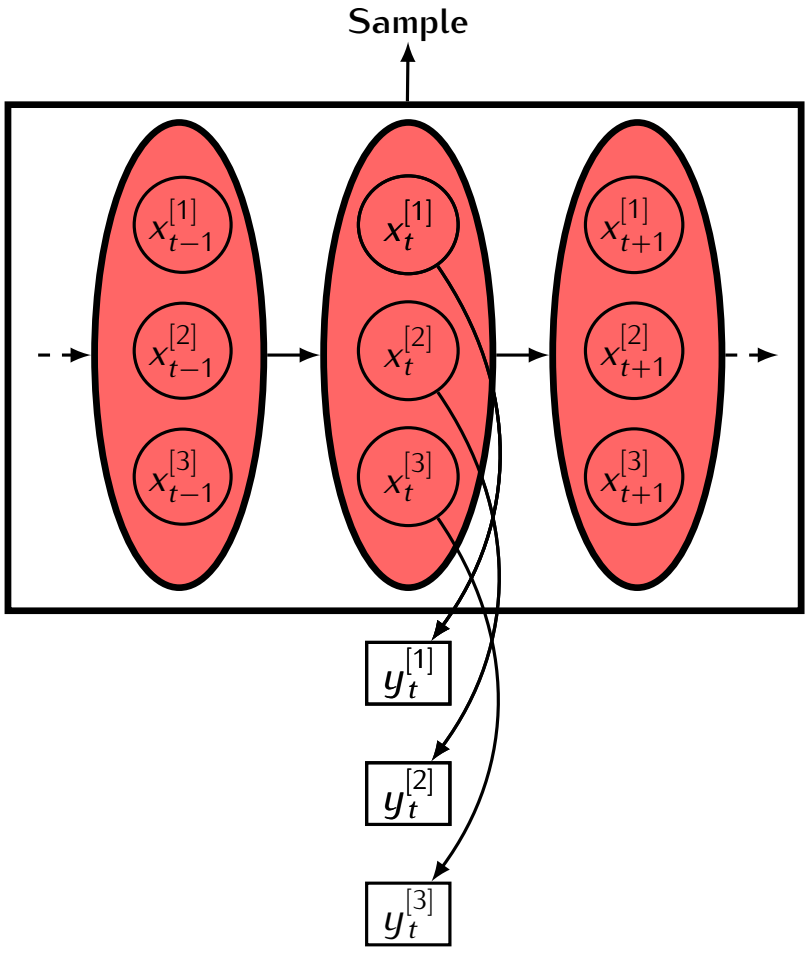
- Block Update Method:
  - ➢ Choose one **block of states** for each individual and propose one of 3 possible changes: **Add** or **Remove** a block of infection/ clearance or **Move** an endpoint of such a block.
- Single-Site Method:
  - ➢ Update each **single node** from its full conditional distribution.
- 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:



- $\boldsymbol{X}_t^{[1:C]} = \left(X_t^{[1]}, X_t^{[2]}, \ldots, X_t^{[C]}\right) \in \mathcal{X}^C = \{1, 2, \ldots, N\}^C$.
- $\left|\boldsymbol{X}_t^{[1:C]}\right| = N^C$.
- Update the whole hidden process $\boldsymbol{X}$ from its full conditional:

$$\boldsymbol{X} \sim \pi\left(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\theta}\right);$$

- Computational complexity: $\mathcal{O}(TN^{2C})$.

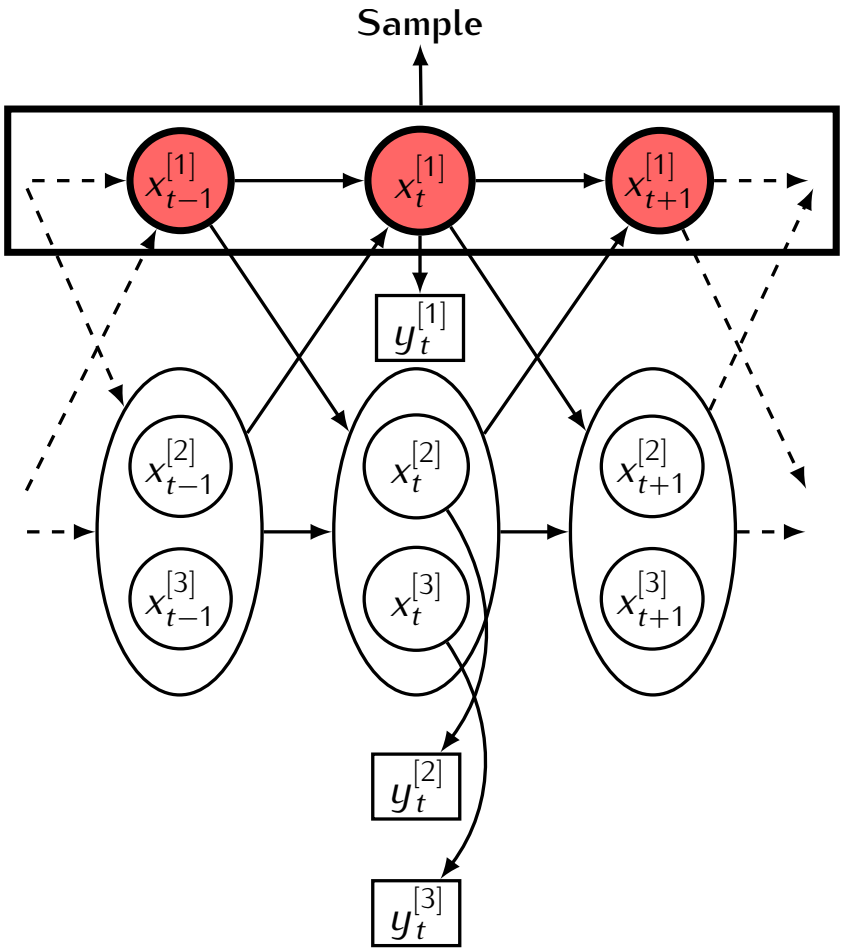
$N =$ number of infection states.

$C =$ number of individuals.

$T =$ number of time-points.

Proposed method: individual FFBS (iFFBS)[1]

Reformulate graph:



- Modification of FFBS.
- Update one individual at a time by sampling from the full conditional:

$$\pi\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right).$$

- Computational complexity reduced to $\mathcal{O}(TCN^3)$.

N = number of infection states.

C = number of individuals.

T = number of time-points.

[1] P. Touloupou, B. Finkenstädt, and S. E. F. Spencer. "Scalable Bayesian inference for coupled hidden Markov and semi-Markov models". In: *Journal of Computational and Graphical Statistics* 29 (2020).

Simulation studies

Application: SIS Markov model

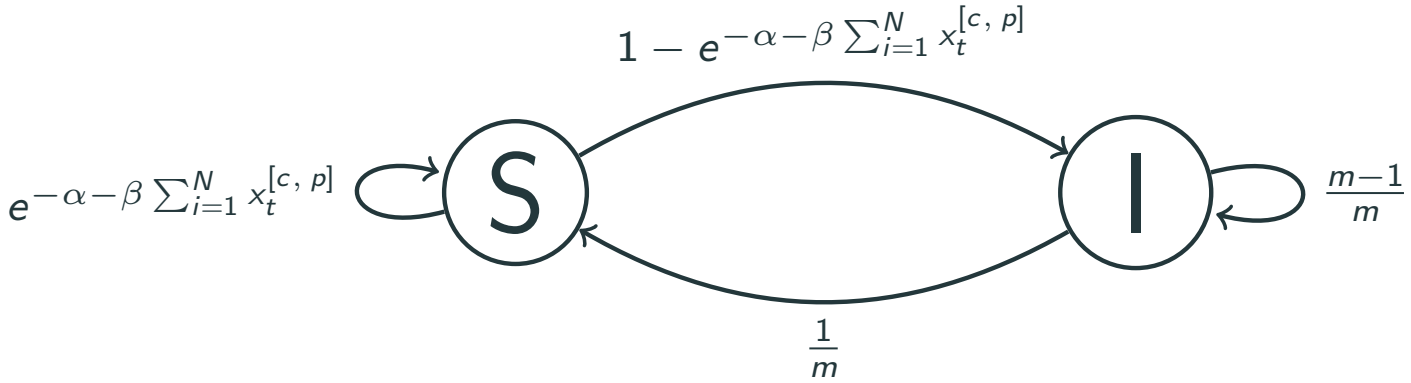
- Stochastic SIS (Susceptible-Infected-Susceptible) transmission model in discrete time.

- $X_t^{[c,\,p]}$ 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.

Application: SIS Markov model

- The transition probabilities between the states are given by:

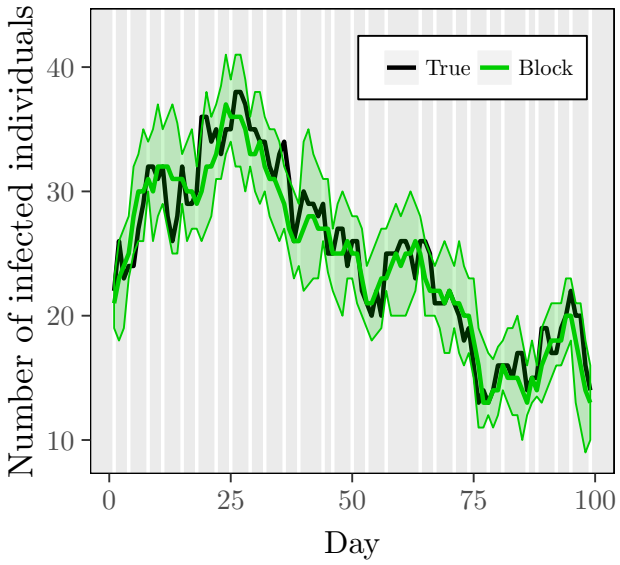


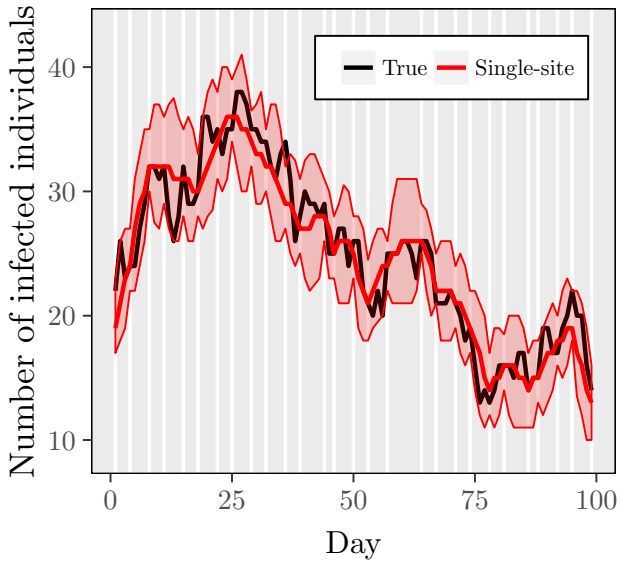
  where α and β are the external and within-group transmission rates, respectively, and m is the mean infection period.

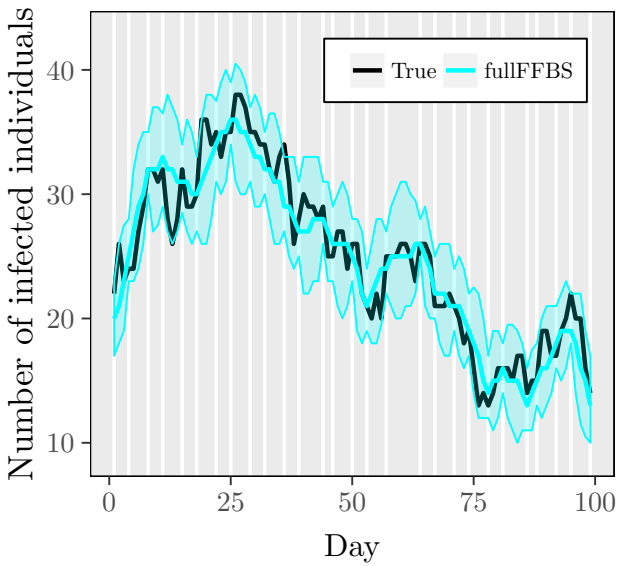
- Individuals are initially infected with probability ν.

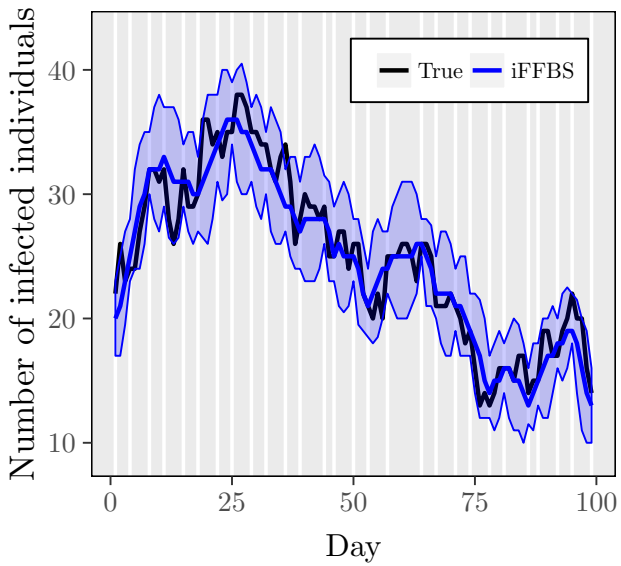
- Tests are assumed to have perfect specificity but imperfect sensitivity.

Comparison of methods: Estimation

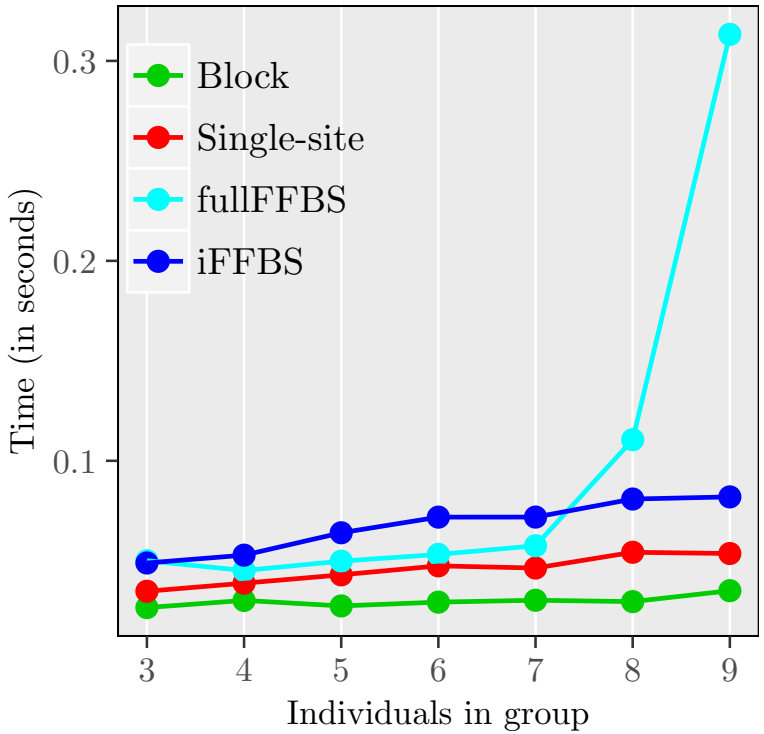


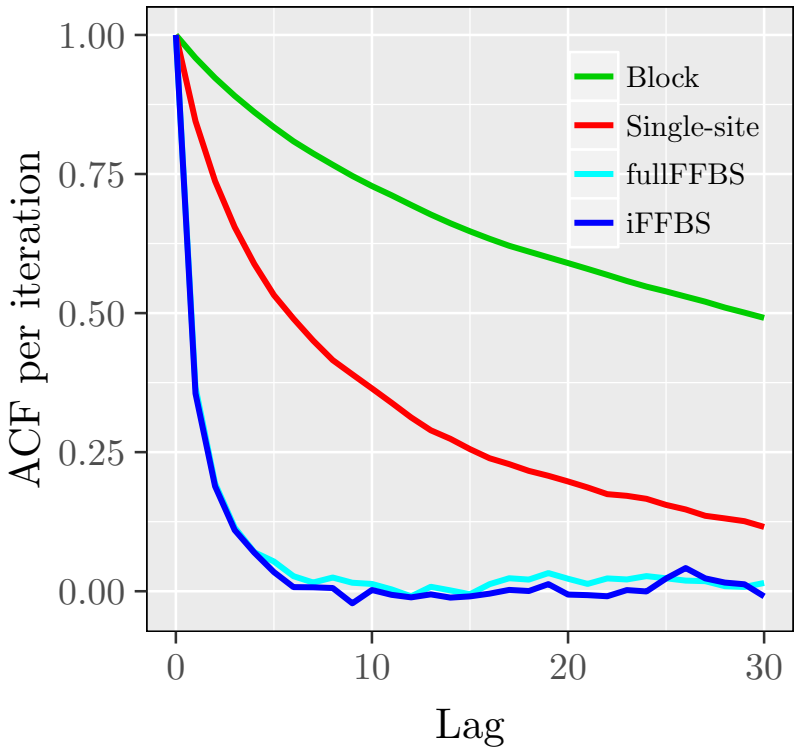




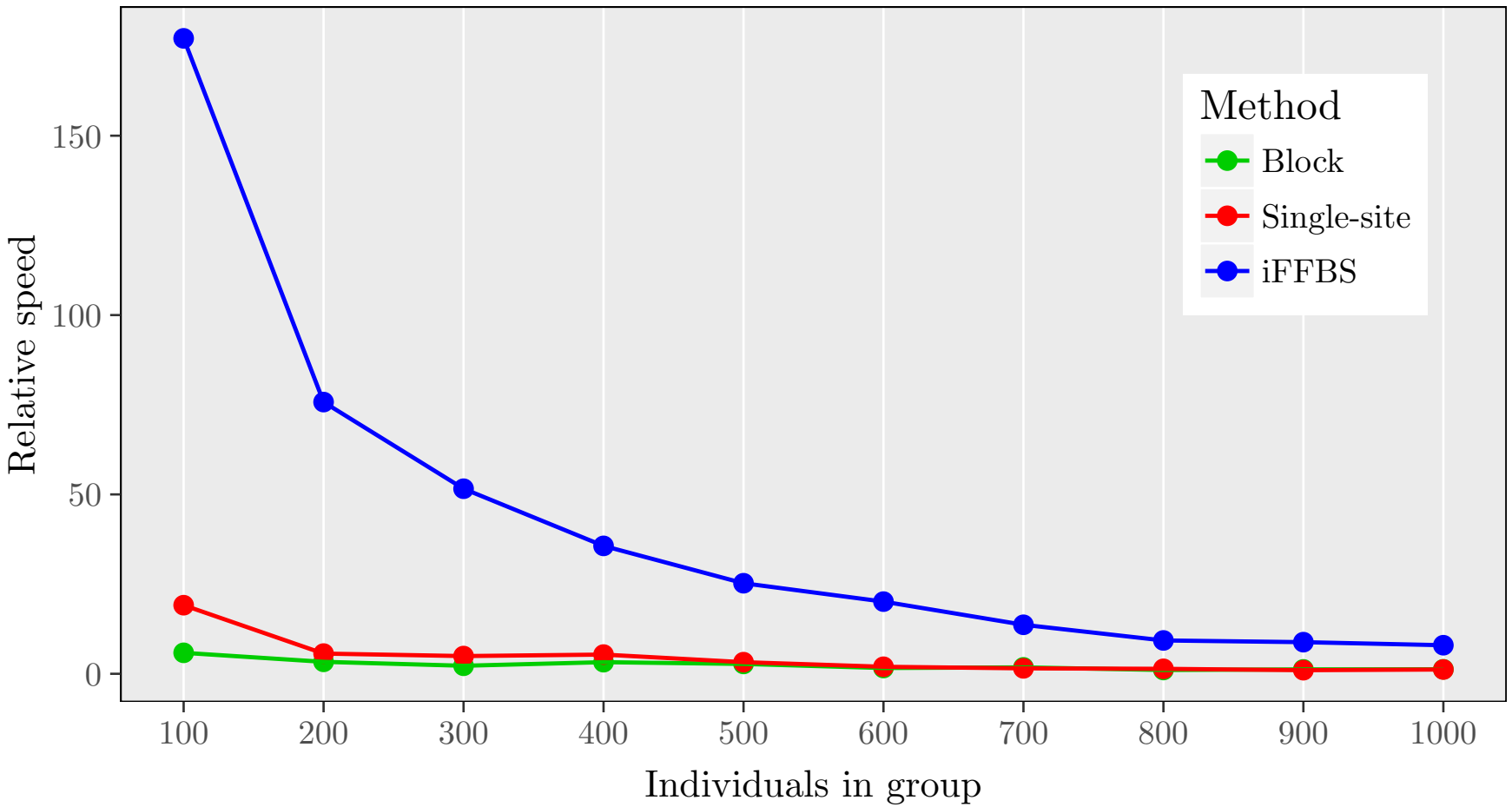


Comparison of methods: Time and ACF





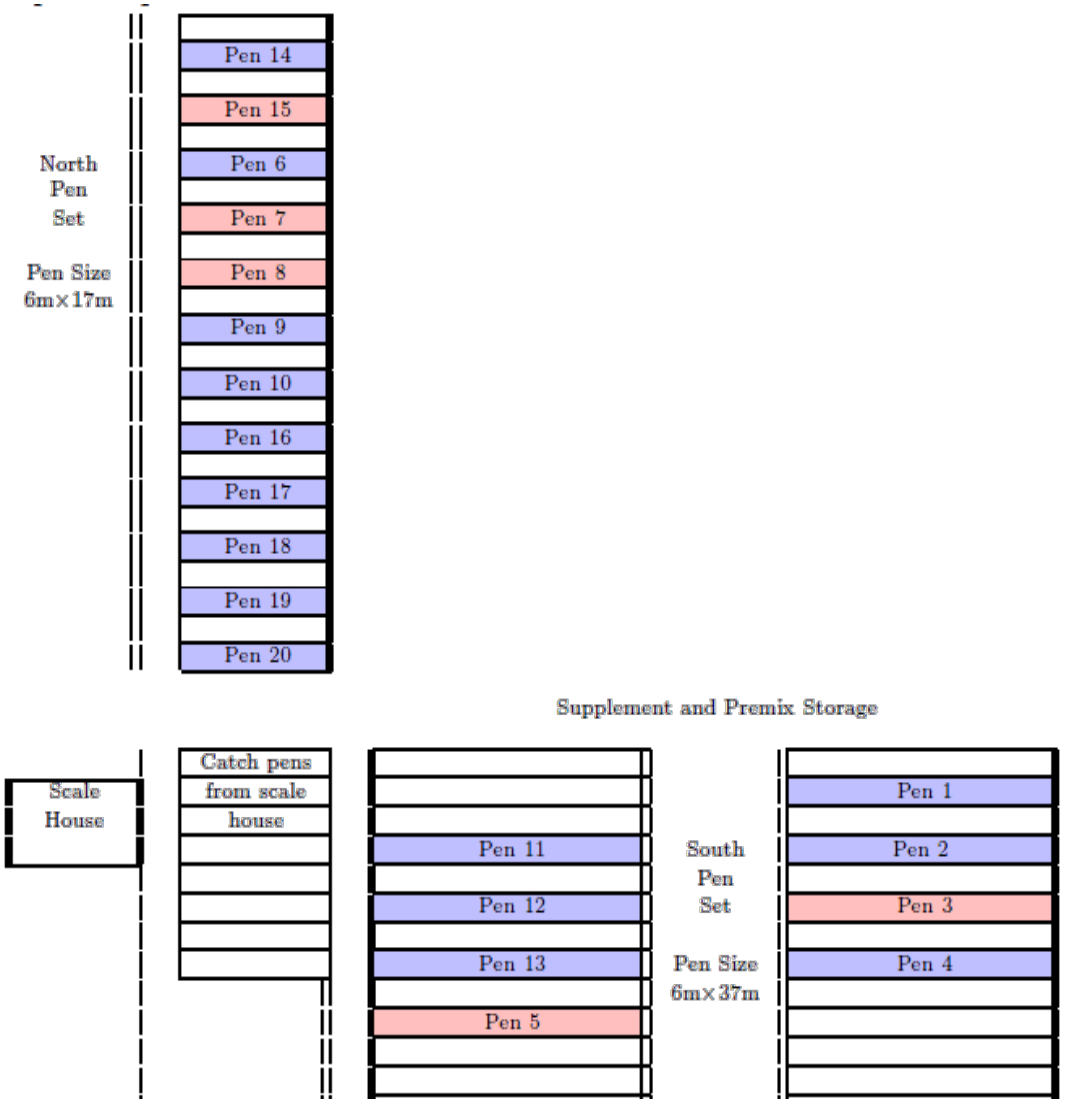
Comparison of methods: Larger population



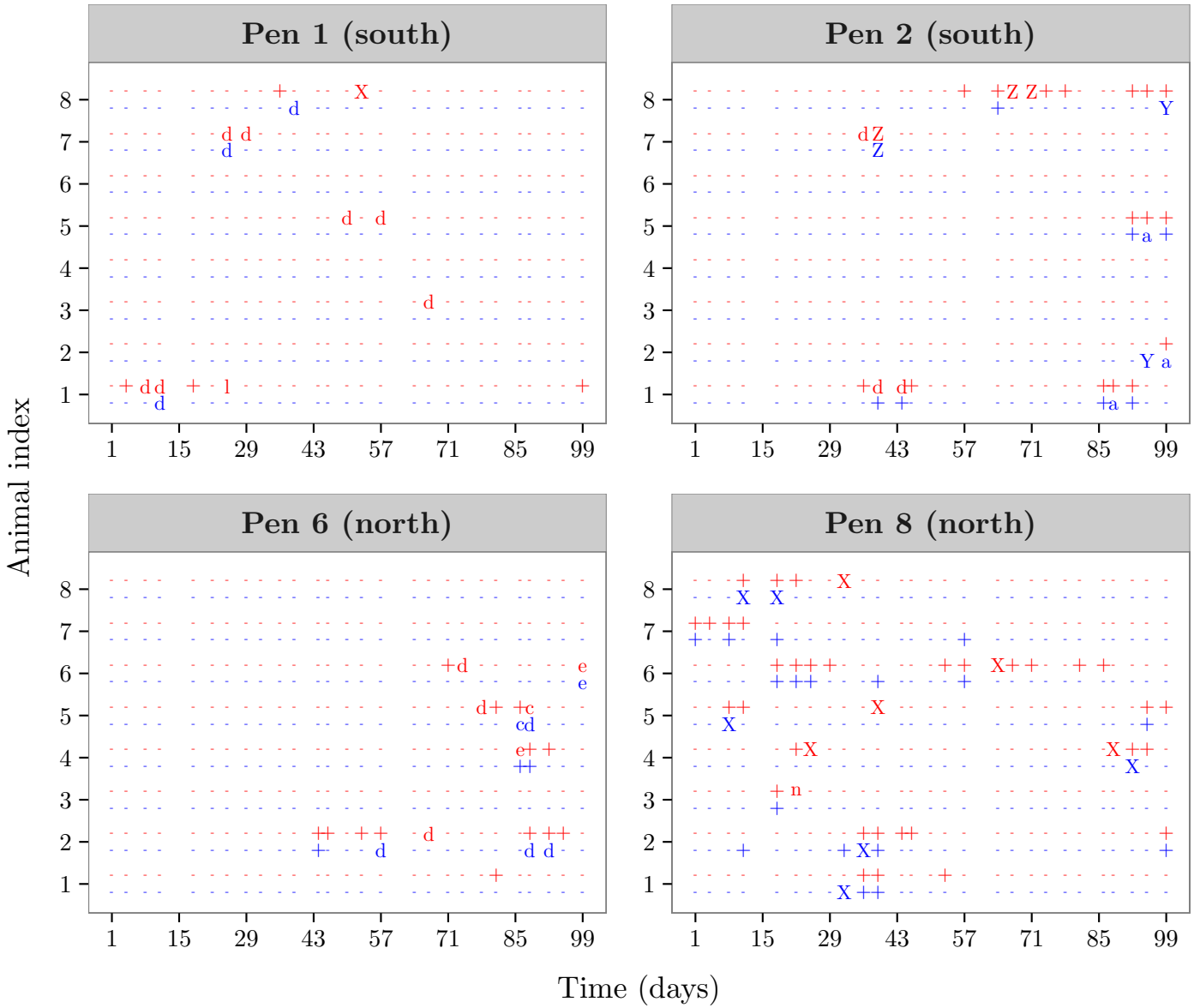
Epidemics with genetic typing data

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.

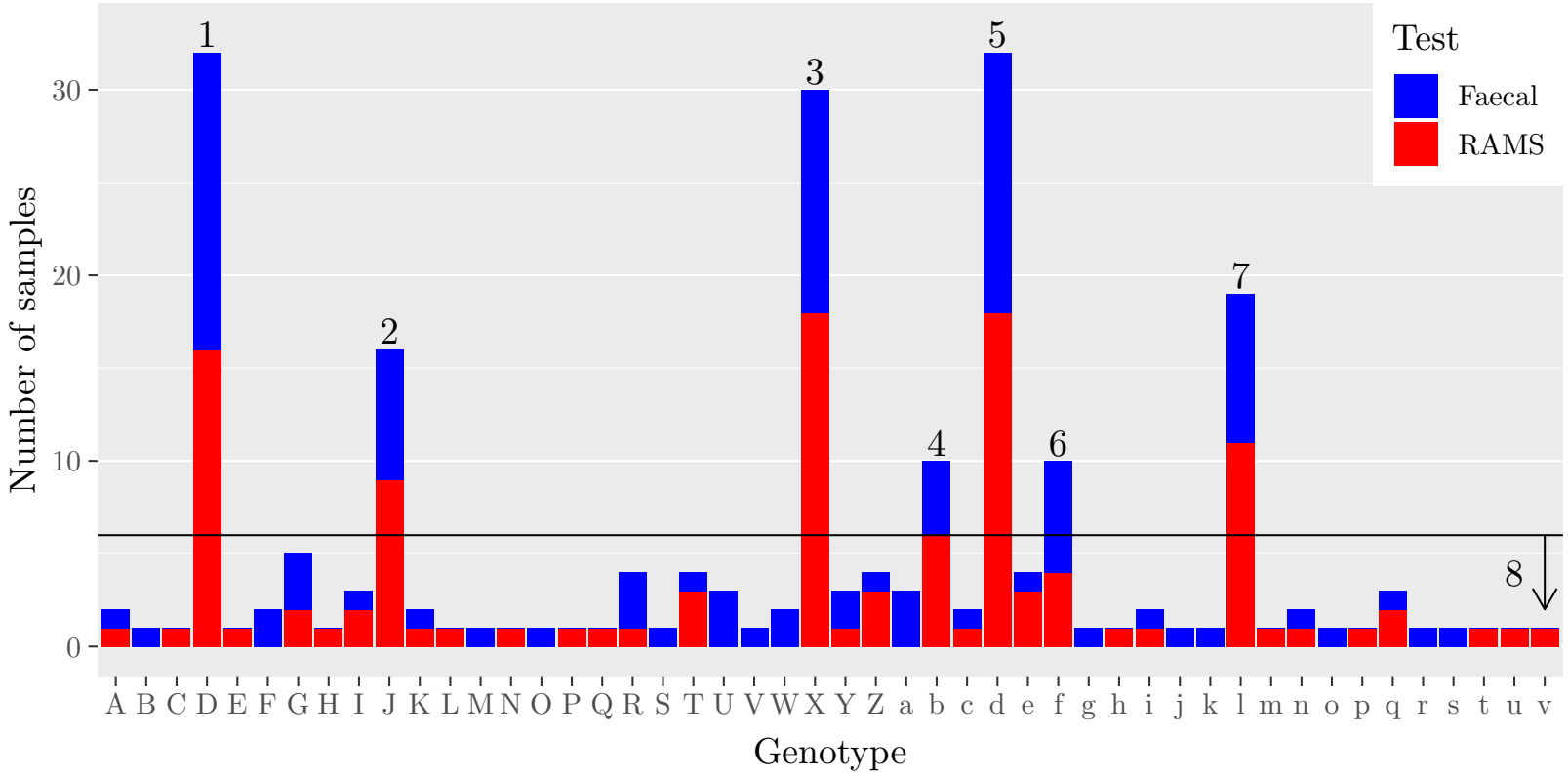


Multi-strain data



Multi-strain data: Summary

- 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[2]

- Stochastic multi-state model in discrete time.

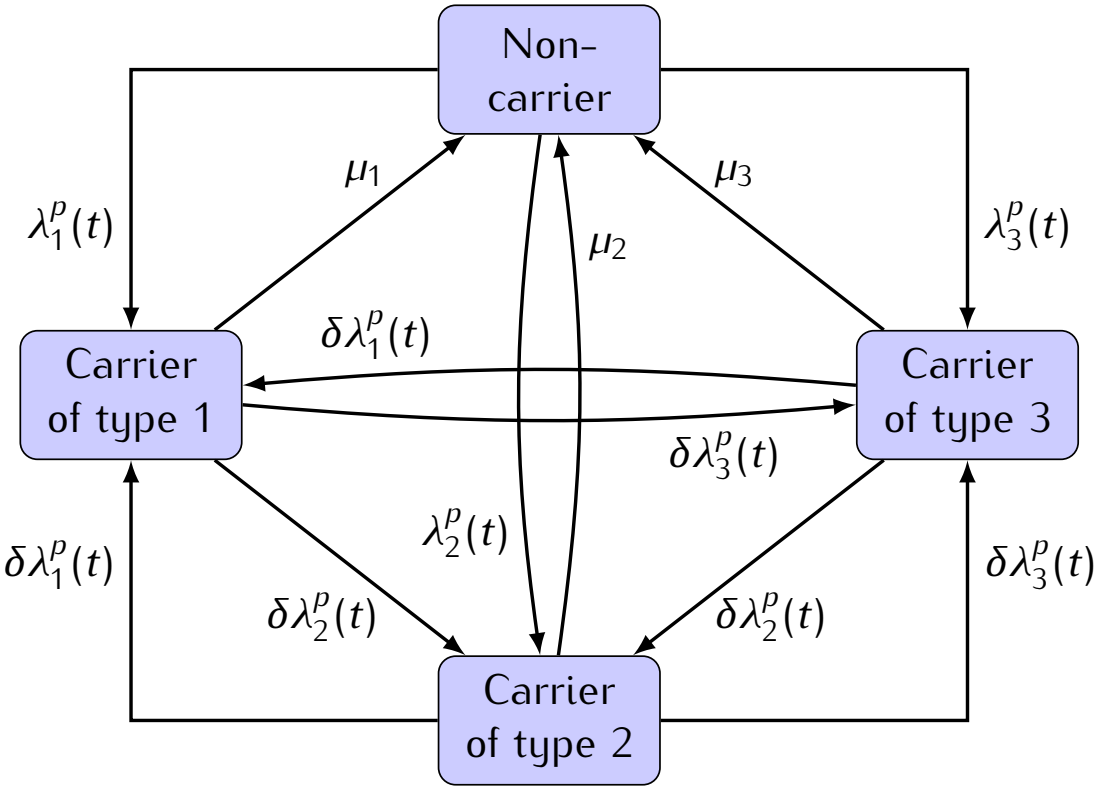
- $X_t^{[c,p]}$ unobserved carriage status for animal c in pen p on day t.
  - ➢ $X_t^{[c,p]} = 0$: non-carrier.
  - ➢ $X_t^{[c,p]} = s$, $s = 1, 2, \ldots, 7$: carriage of one of the common genotypes.
  - ➢ $X_t^{[c,p]} = 8$: carriage of the remaining genotypes (pooled group).

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

[2] P. Touloupou et al. "Bayesian inference for multi-strain epidemics with application to *Escherichia coli* O157:H7 in feedlot cattle". In: *The Annals of Applied Statistics* (in press).

Transitions between the states

- Acquisition rate: $\lambda_s^p(t) = \alpha_s + \beta_s \sum_{i=1}^{C} \mathbf{1}_{\{X_t^{[c,p]}=s\}}$
- Clearance rate: μ_s
- Relative colonisation rate in a carrier versus non-carrier: δ



Example of an epidemic model with 3 competing types.

Comparing parameters between genetic types

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

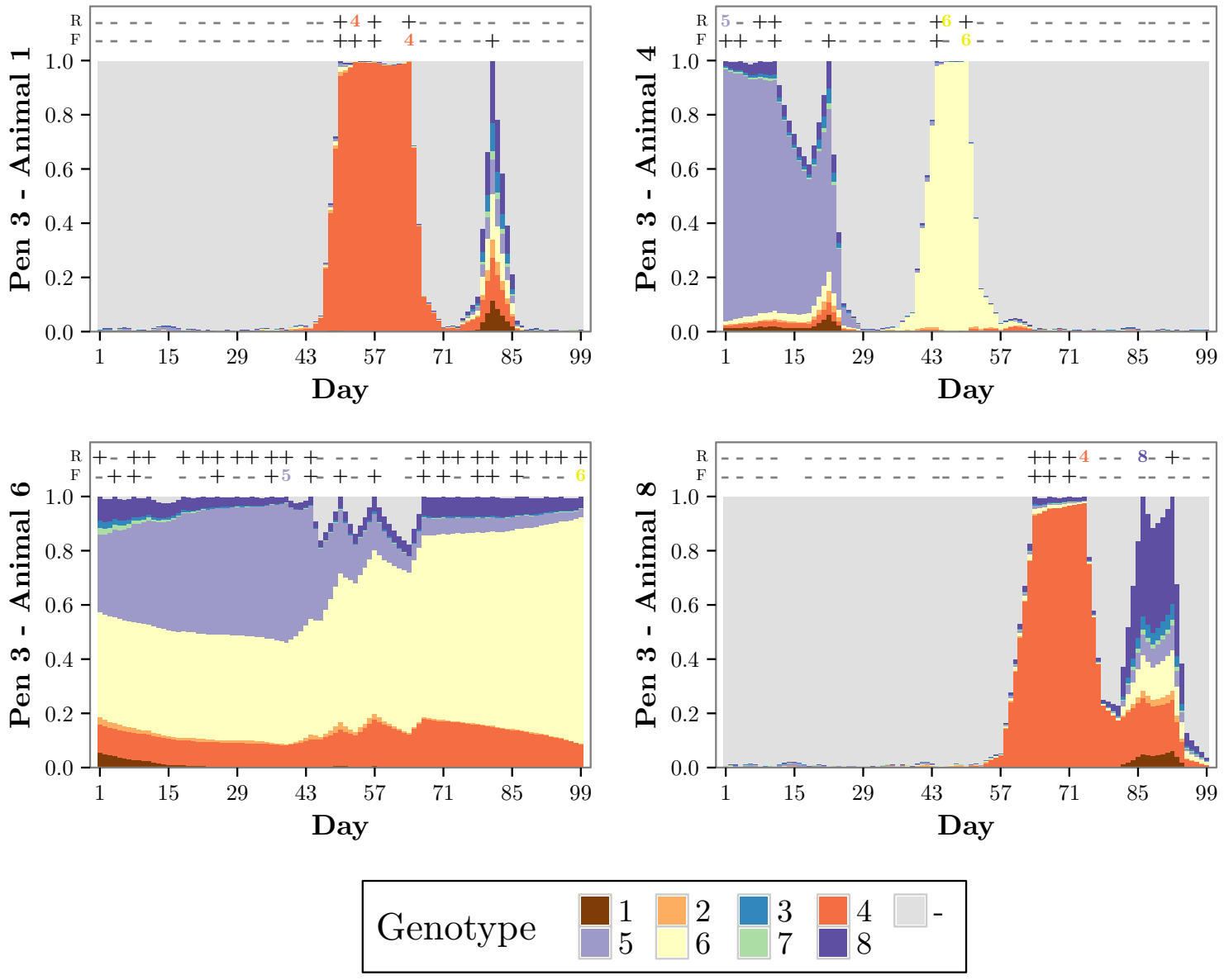
Comparing parameters between genetic types

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

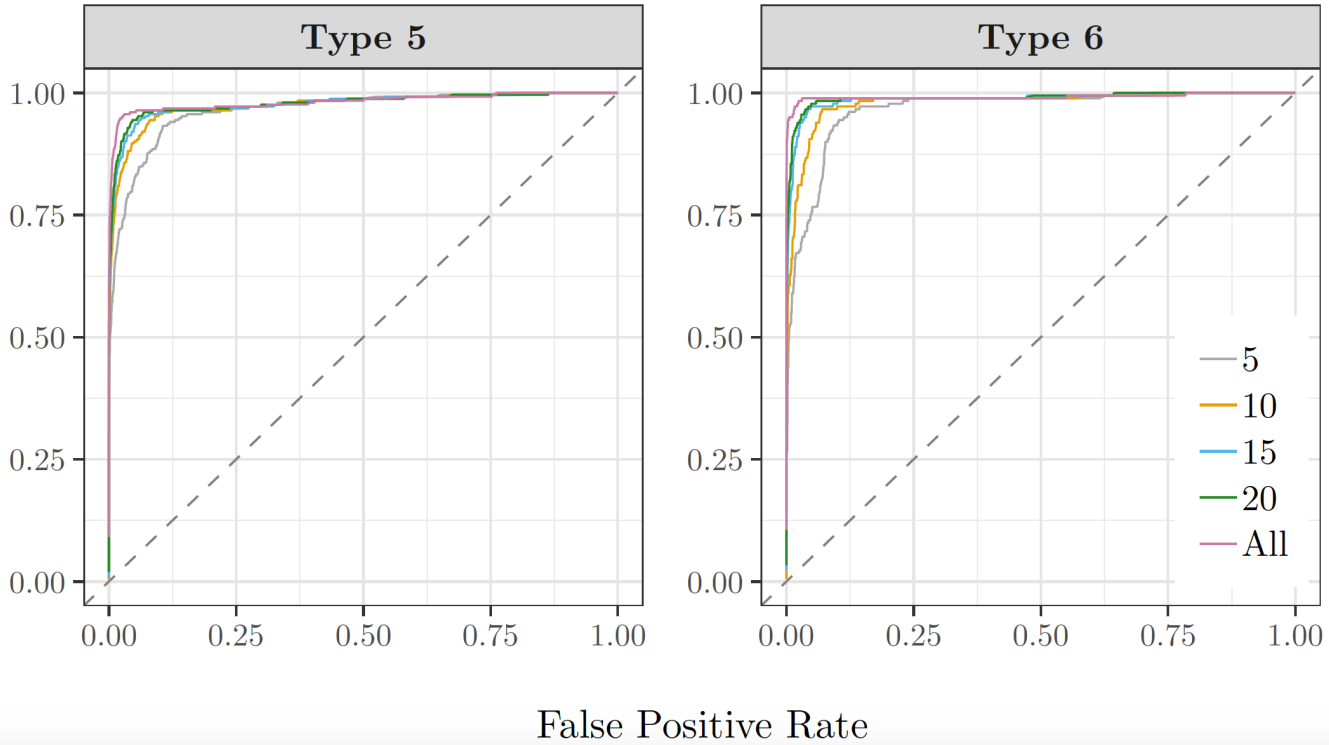
Rest of the parameters

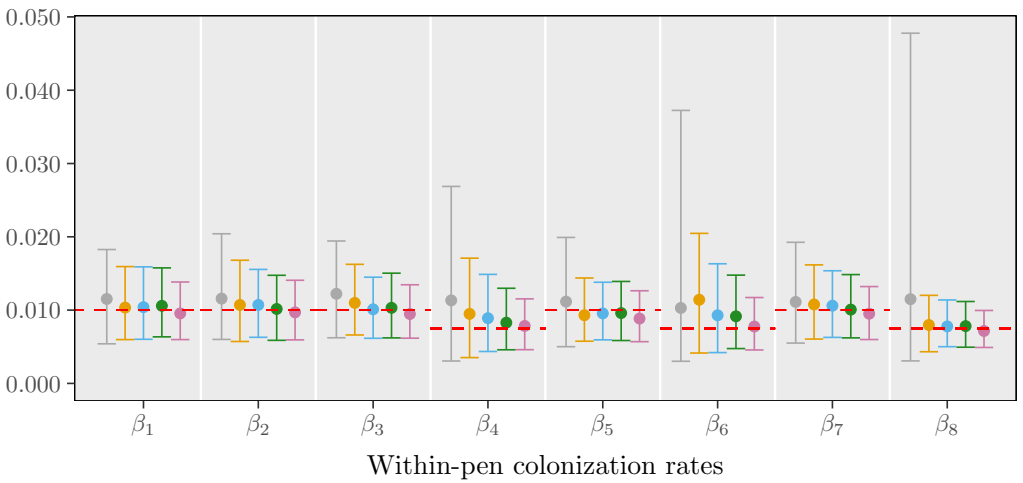
- The median relative colonisation rate in a carrier versus non-carrier individual is 0.842.
- Test sensitivities:
  - RAMS test: 76%,
  - Feacal test: 46%.
- 81.6% of the common genotypes are correctly classified as the right type.
- 1.2% are misclassified as another common type.
- 17.2% are misclassified 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





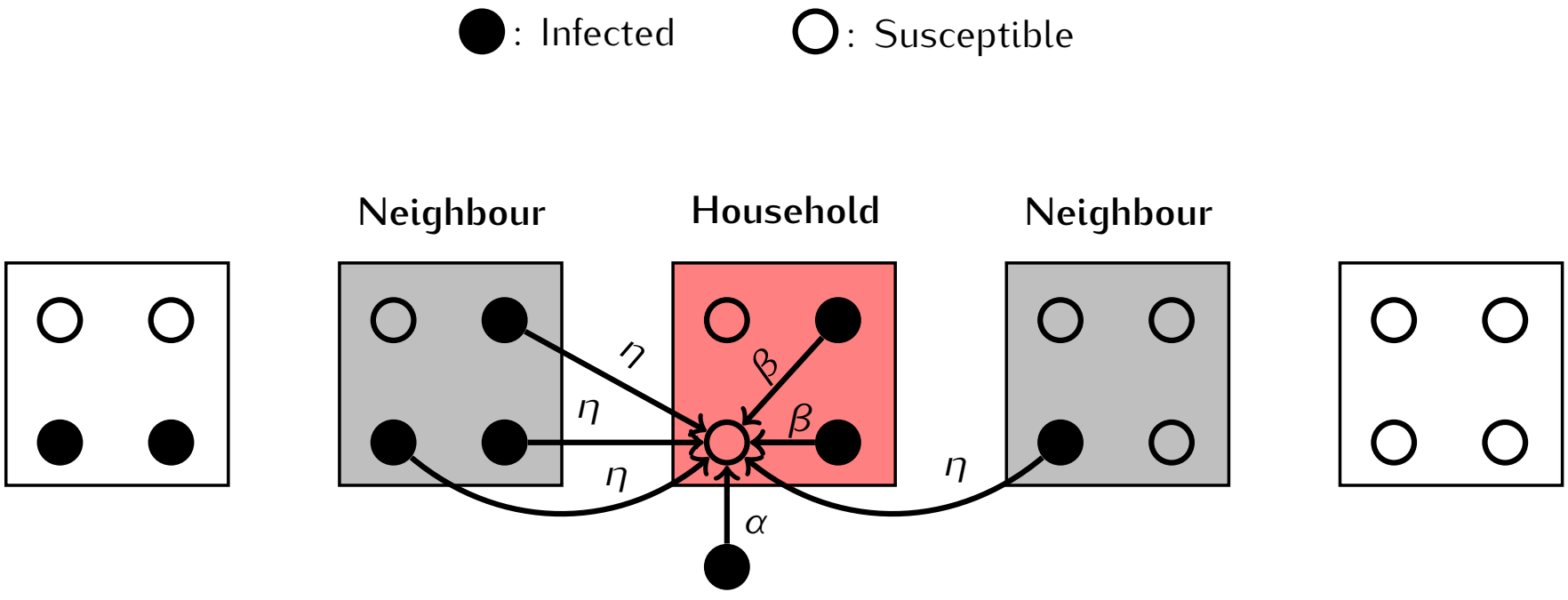
Discussion

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*[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).

[3] Panayiota Touloupou and Simon E. F. Spencer. *epiPOMS: Bayesian Inference for Partially Observed Multi-Strain Epidemics*. R package, version 0.1.0. 2020.

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

- Carter, C. K. and R. Kohn. "On Gibbs Sampling for State Space Models". In: *Biometrika* 3 (1994).
- Dong, W., A. Pentland, and K. A. Heller. "Graph-coupled HMMs for modeling the spread of infection". In: *arXiv preprint arXiv:1210.4864* (2012).
- Spencer, S. E. F. 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).
- Touloupou, P., B. Finkenstädt, and S. E. F. Spencer. "Scalable Bayesian inference for coupled hidden Markov and semi-Markov models". In: *Journal of Computational and Graphical Statistics* 29 (2020).
- Touloupou, Panayiota and Simon E. F. Spencer. *epiPOMS: Bayesian Inference for Partially Observed Multi-Strain Epidemics*. R package, version 0.1.0. 2020.
- Touloupou, P. et al. "Bayesian inference for multi-strain epidemics with application to *Escherichia coli* O157:H7 in feedlot cattle". In: *The*

Misclassification Matrices

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

$$\mathbf{E}^{R_+} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ n_s \end{array} \begin{array}{c} \begin{array}{cc} 0 & + \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1-\theta_R & \theta_R \\ \vdots & \vdots \\ 1-\theta_R & \theta_R \end{bmatrix} \end{array}$$

where θ_R is the sensitivity of the RAMS test and is denoted by $\theta_R = \mathbb{P}\left(R_t^{[c,\,p]} = + \mid X_t^{[c,\,p]} = r\right)$.

Misclassification matrices

For a positive sample that was genotyped we introduce additional parameters θ_C, θ_S and θ_U:

$$\mathbf{E}^{R_s} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ n_s - 1 \\ n_s \text{ (Type U)} \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & \cdots & \cdots & \cdots & n_s - 1 & n_s \text{ (Type U)} \end{array} \\ \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1-\theta_R & \theta_C\,\theta_R & \frac{\theta_S\,\theta_R}{n_s-2} & \cdots & \cdots & \frac{\theta_S\,\theta_R}{n_s-2} & (1-\theta_C-\theta_S)\,\theta_R \\ \vdots & \frac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & \frac{\theta_S\,\theta_R}{n_s-2} & \cdots & \frac{\theta_S\,\theta_R}{n_s-2} & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \frac{\theta_S\,\theta_R}{n_s-2} & \cdots & \frac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & \frac{\theta_S\,\theta_R}{n_s-2} & \vdots \\ \frac{\theta_S\,\theta_R}{n_s-2} & \cdots & \cdots & \frac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & (1-\theta_C-\theta_S)\,\theta_R \\ 1-\theta_R & \frac{\theta_U\,\theta_R}{n_s-1} & \cdots & \cdots & \cdots & \frac{\theta_U\,\theta_R}{n_s-1} & (1-\theta_U)\,\theta_R \end{bmatrix} \end{array}$$

such that, for all $r \neq 0$, the probabilities
$e^{R_s}_{r,0} = \mathbb{P}\left(R_t^{[c,p]} = 0 \mid X_t^{[c,p]} = r\right) = 1 - \theta_R$ and $\sum_{s=1}^{n_s} e^{R_s}_{r,s} = \theta_R$.