

Modelling serological data using reversible jump Markov chain Monte Carlo

David Hodgson, James Hay(?) and Adam Kucharski

Center of Mathematical Modelling of Infectious Diseases

London School of Hygiene and Tropical Medicine

January 24, 2024

1 Overview of serological modelling

1.1 Introduction

Serological samples can be analysed to detect the presence of biomarkers made in response to an infection long after the infection has cleared.[?] Therefore, analysing serological samples allows researchers and healthcare professionals to deduce crucial information about the epidemiology of a pathogen at the individual and population level, which active virological surveillance systems may otherwise miss.

On the individual level, after measuring antibodies to a specific pathogen, infection is usually inferred using either i) an antibody threshold level (seropositive) or ii) a threshold fold-rise between a pair of samples (seroconverted).[?] Often, researchers are interested in understanding how seropositivity and seroconversion rates change according to controlled host factors, such as age, XXXX etc.[?] On the population level, serological samples which are representative of a population (e.g. cross-sectional samples) can be used to estimate the prevalence of infectious diseases (seroprevalence) and determine how seroprevalence changes over time according to host factors.[?] Estimates of seropositivity, seroconversion, and seroprevalence can help the understanding of the immune system's ability to combat various pathogens, aid in developing new targeted intervention programmes and provide insights into the transmission dynamics of infectious diseases. The methods used to analyse serological samples to inform epidemiology and public health have been termed 'serodynamics' and have recently been reviewed. [?]

Serological samples play an increasingly important role in public health efforts to combat and control infectious diseases.[?] However, inferring infection through seropositivity or seroconversion requires deriving an absolute or relative threshold hold value, and these are often determined by rule-of-thumb heuristics (e.g. flu, 4-fold-rise for conversion, titre of 1:40 HAI for seropositivity).[?] However, antibody responses vary greatly between individuals for many pathogens; therefore, relying on these heuristics to determine infections in serology studies can lead to incorrect infection status being inferred, potentially leading to biased estimates.[?] Consequently, a better understanding of the kinetics of antibody trajectories post-vaccination and infection can help establish and ascertain the accuracy of existing heuristics and be used to more accurately infer infection status.

1.2 Antibody kinetics

Modeling antibody kinetics involves using mathematical and statistical techniques to simulate the trajectories of antibodies in response to an infection or vaccination. [?] Typically, this involves the use of mathematical equations and statistical methods to describe the time-dependent changes in antibody levels within an individual or a population. This process is essential for understanding how antibody levels evolve and, therefore, potentially provide protection against infectious diseases. Various functional forms

have been used to model the individual-level kinetics of antibody trajectories,[?] typically it follows a three-stage process:

- *Initial Response*: The trajectories often start by capturing the initial antibody response to a pathogen or vaccine. This phase is characterized by a rapid increase in antibody levels as the immune system recognizes and mounts a defence against the antigen.
- *Peak Antibody Level*: The trajectories then rise to a peak antibody level, which is the highest concentration of antibodies reached during the immune response. This peak can vary depending on factors like the strength of the immune response and the nature of the antigen.
- *Decay Phase*: After the peak, there is a decline in antibody levels. Antibodies have a finite lifespan in the bloodstream, and their concentration gradually decreases as the pathogen is cleared or the vaccine antigen wanes. Often antibodies secreted from now-establish long-lived plasma cells, the decay phase trajectories fall to a set point titre.[?]

Modeling antibody kinetics provides several important benefits. First, by understanding the rate of antibody decline, models can estimate how long an individual's immunity is likely to last after infection or vaccination. This information is critical for designing vaccination schedules and determining the need for booster shots and permits the creation of better heuristics for determining infections using serology. Kinetics are also useful for optimizing vaccination strategies and can help identify the optimal timing and frequency of booster vaccinations to maintain protective antibody levels within a population. This is especially important for vaccine-preventable diseases with varying immunity levels such as influenza and COVID-19.

1.3 Correlates of protection

A correlate of protection is an immune function or biomarker that correlates with and or may be biologically responsible for protection against infection or disease. Correlates of protection have been established for influenza, Hepatitis A and B, Measles, Polio, Rabies, Yellow fever and more.[?] Correlates of protection help researchers understand the specific immune responses needed to prevent or control an infectious disease. This knowledge is pivotal in designing and optimizing vaccines to induce the required immune response effectively.[?]

Correlates of protection can be established by i) comparing the immune response of those protected by the vaccine with so-called 'breakthrough cases', where clinical disease manifests despite prior vaccination and ii) human challenge studies.[?]

A natural biomarker for a correlate of protection from infection is the amount of neutralising antibodies in the serum as this measures a serum's ability to prevent viral particles from infecting vulnerable cells. Therefore, those with high levels of neutralising antibodies could abort an infection by neutralising viral particles in-host even if exposed to a virus. Determining the immunological profiles of individuals are exposed to an infection but manage to abort it due is crucial for determining a universal correlate of protection. Usually, these individuals are identified through challenge studies, allowing for a direct evaluation of correlates of protection, however these studies are expensive, difficult to run, can only be performed on healthy individuals, and are only possible for pathogens with general pathogenesis.

Using serological samples to establish a correlate of protection would solve the aforementioned problems with challenge studies as they are cheaper to conduct and less invasive. However, determining correlates of protection through serological studies is challenging to establish because those who have been exposed but then experience an abortive infection generally leave no measurable antibody imprint. Therefore serological studies can be augmented with either immunological profiling to determine other immunological biomarkers that indicate abortive infection[?] or include intensive contact tracing with the serological studies to determine exposure rates between individuals[?]. Augmented serological studies with these additional increase the complexity and cost of a study and, therefore, are not feasible in many settings.

1.4 Overview of modelling framework

If antibody trajectories are known for every individual throughout a study and the time of infection is known, then we can determine the probability of infection for a given titre value. Though not strictly a correlate of protection, as we do not know whether non-infected individuals have been exposed, this can be useful for determining associations between infected and non-infected individuals for a given population. However, relying on this sample to establish correlates of protection can lead to bias estimates XXX.

In this document, we present a single modelling framework which takes individual-level serological sample data and uses changes in antibody titres over time to determine i) which individuals have seroconverted throughout the study, ii) subsequent antibody kinetics of these infection individuals, iii) which individuals are exposed throughout the study, and iv) the correlate of protection preventing exposed individuals from becoming infected. Though the infection and exposure status is often unknown for most individuals in a serological study, we find that by using broad, biologically-informed mechanistic forms for the antibody kinetics and correlation of protection, the infection and exposure status of individuals and the population are recoverable through the interdependencies of mechanisms i)–iv) within a Bayesian probabilistic framework.

2 Simulated serological data from serosim

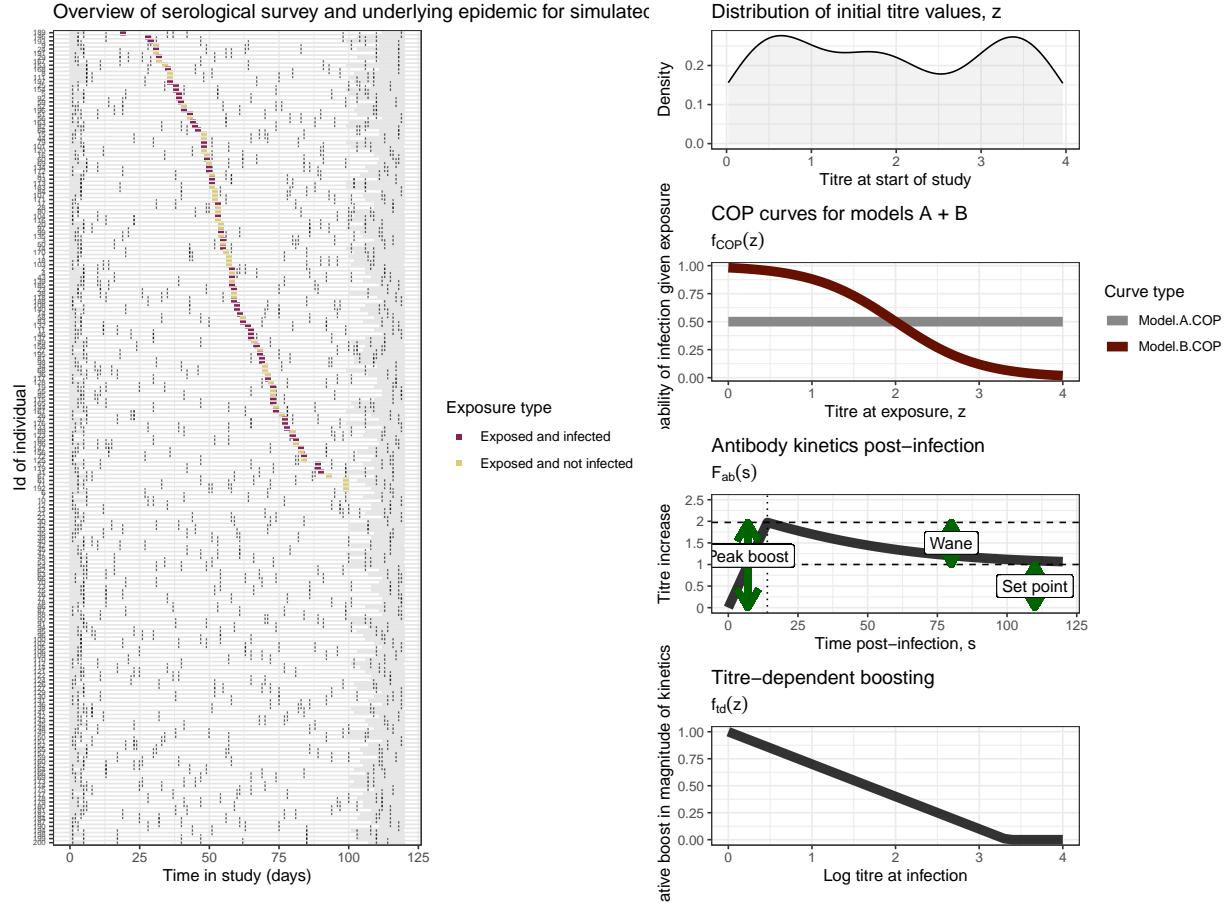


Figure 1: Schematics showing the simulated data structure from `serosim`[ref]

To demonstrate the effectiveness of our modelling framework, we simulate serological data using the `serosim`[ref] R package to demonstrate its capabilities at simulation recovery. We simulate continuous epidemic serosurveillance (CES) cohort data, which represents a study in which individuals are followed over a period spanning an epidemic wave and bled at multiple random time points throughout. The simulated data includes $M = 200$ individuals with serological samples taken within the first seven days of the study's starting and a sample within the last seven days of the study's ending. These individuals also had three samples taken randomly throughout the study (over the 120-day epidemic wave). Each individual has a 60% chance of exposure to the virus over the study timeframe and can have a maximum of one exposure. To model an even epidemic peak, we simulate the exposure time for each individual from a normal distribution, $N(60, 20)$ days.

We define a correlate of protection as the probability of infection given a titre value at exposure. Two sets of data are simulated with two different correlates of protection; one is uniform at 50% for all titres at exposure (COP model A) and thus represents no titre-dependent protection correlation. The second follows a logistic distribution (COP model B) of the form:

$$f_{cop}(x, \beta_0, \beta_1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))} \quad (1)$$

where $\beta_0 = 2$ and $\beta_1 = 2$ in the simulated data and x is the titre value at exposure. This represents a pathogen for which higher antibody titres are associated with higher levels of protection from infection. Note in this data we assume antibody trajectories remain constant until the timing of infection, such $f_{cop}(X_{j,t}, \beta_0, \beta_1) = f_{cop}(Z_j^0, \beta_0, \beta_1)$ where Z_j^0 is the antibody titre of an individual at their first bleed at the start of the study.

Following infection, the antibody kinetics are assumed to follow a linear rise to a peak at 14 days, followed by an exponential decay to a set-point value as defined in X[ref]. The formula for this biphasic trajectory is given by Equation 2.

$$f_{ab}^1(s, a, b, c) = \begin{cases} \ln(\exp(a) + \exp(b))/14, & \text{if } s \leq 14 \\ \ln(\exp(a) \exp(-(b/10)(t - 14)) + \exp(c)), & \text{if } s > 14 \end{cases} \quad (2)$$

where $a = 1.5$, $b = 2$, and $c = 1$ are values in the simulated data (Figure) and s is the number of days post infection. We also assume that the magnitude of these dynamics depends on pre-existing titre values, with higher pre-existing values seeing attenuated dynamics relative to lower pre-existing titre values. The titre dependent boosting is assumed to follow a linear decay truncated at 0, that is, given a titre value z , $f_{ab}^2(Z_j^0, \alpha) = \max(1 - \alpha z, 0)$. where $\alpha = 0.3$ in the simulated data. Therefore, the model estimated titre value at time t for individual j , given time s since infection with a titre value at infection of Z_j^0 , is given by;

$$f_{ab}(s, Z_j^0, a, b, c, \alpha) = f_{ab}^1(s, a, b, c) f_{ab}^2(Z_j^0, \alpha) \quad (3)$$

To model heterogeneity in individual-level antibody kinetics, we simulate a, b, c , and α from normal distributions where the mean μ is their simulated values and the standard deviation given by $\mu\sigma^*$. We simulate three levels of uncertainty for $\sigma^* = \{0.1, 0.3, 0.5\}$, for both of the COP models, giving six simulated datasets in total. A schematic showing how these levels of uncertainty influence the variability of the antibody kinetics trajectories is shown in Figure 2. We will fit the to-be-described methods to these datasets and explore how the correlation of protection and level of variability in antibody kinetics impacts the framework's ability to recover simulated data.

We note that the simulated data for the correlate of the protection doesn't not exact match the functional form chosen when it is inferred from the infected population ??

Models with different levels on individual variability in antibody kinetics

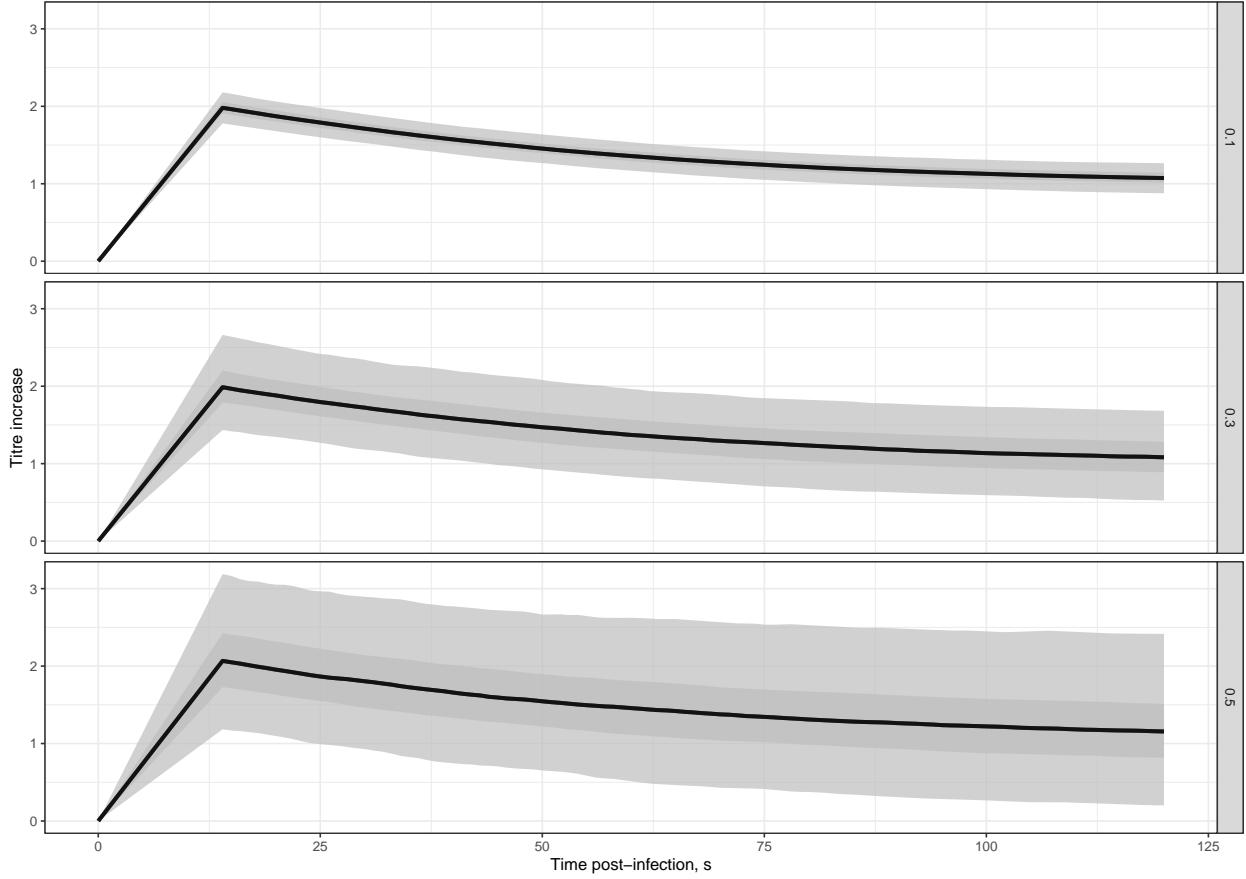


Figure 2: Schematics showing three levels of individual-level uncertainty and the impact on the variability of antibody kinetics.

3 Inference with known exposure status

Before we look at the reversible-jump mcmc algorithm (RJMCMC), we will show simulation recovery in a simplified framework where we assume the exposure status of every individual (represented by vector \mathbf{E}) and exposure time (represented by vector \mathbf{E}^τ) is known. Though knowing this information is rarely feasible in practice, working through this example will help explain how the inference on the fitted parameters θ and infection state \mathbf{I} work without needing to describe the more complex inference using RJMCMC.

3.1 Mathematical representation of framework

Let the binary vector $\mathbf{E} = \{E_1, E_2, \dots, E_M\}$, represent the exposure status of each individual j where $E_j = 0$ is not exposed and $E_j = 1$ is exposed and let $n_{\mathbf{E}} = \sum_{j=1}^M E_j$ be the number of exposed individuals. Then, let the vectors $\mathbf{E}^\tau = \{E_1^\tau, E_2^\tau, \dots, E_{n_{\mathbf{E}}}^\tau\}$ and $\mathbf{I} = \{I_1, I_2, \dots, I_{n_{\mathbf{E}}}\}$ be the timing of the exposure and the infection state respectively for each individual which is exposed. The infection state is a binary vector where $I_j = 0$ is not infected, and $I_j = 1$ is infected. Let $Z_{j,t} \in \mathbf{Z}$ represent the dataset of titre values for individual j and at time t .

We define several functions to help us calculate the likelihood of our model. First, we assume that the model predicted antibody titre at time t in the study ($X_{j,t}$) can be derived given the infection status I_j and timing of exposure E_j^τ . If a person is not infected, their starting titre value (Z_i^0) remains unchanged from

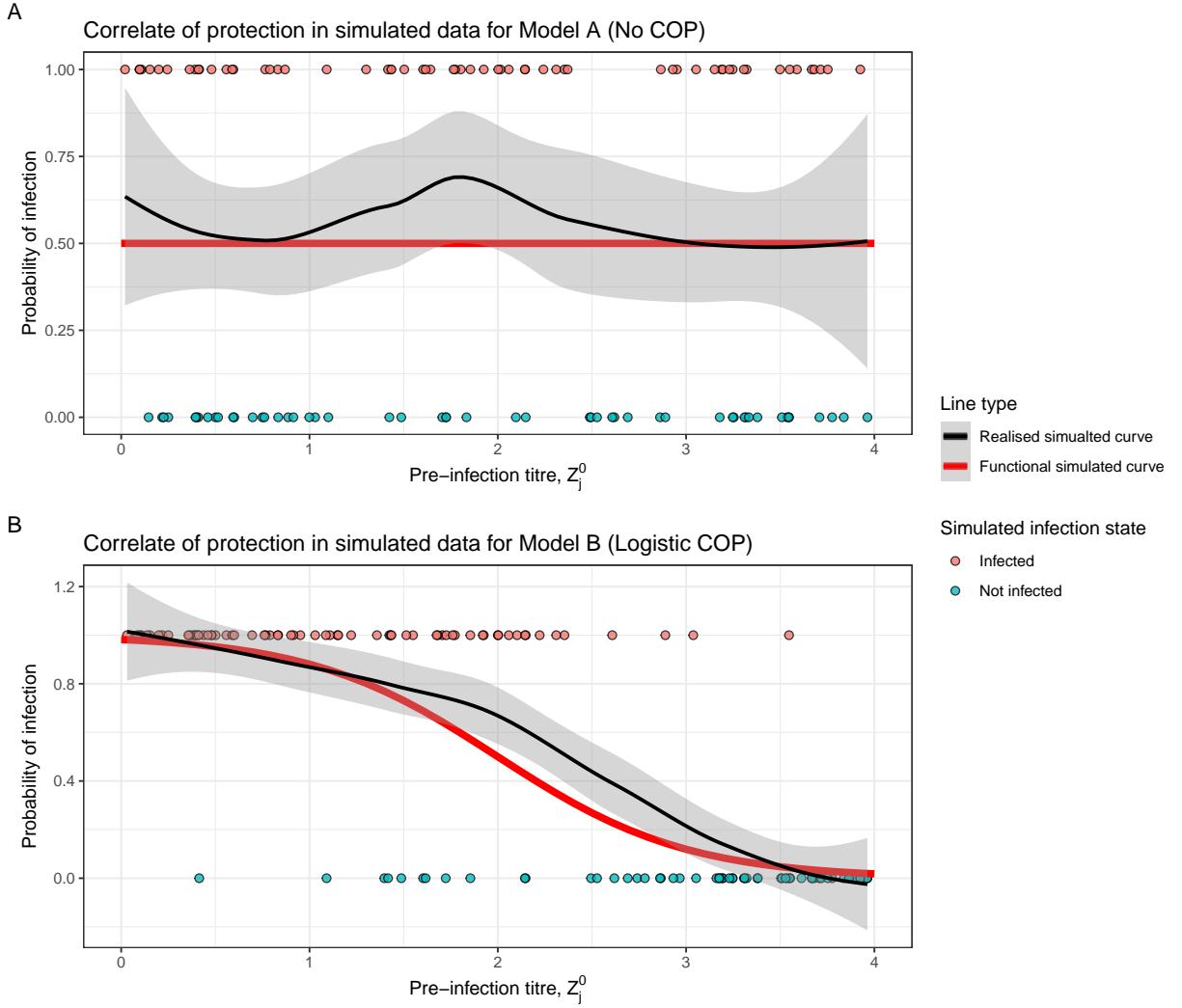


Figure 3: Schematics showing the difference between the functional form chosen to simulate the COP and the recovered COP from the exposed individuals.

the start of the study. If the person is infected, their titre remains unchanged until the point of infection, at which point they follow the dynamics highlighted in Equation 3. The deterministic function for calculating $X_{j,t}$ value is given by

$$X_{j,t} = F_{ab}(I_j, E_j^\tau, \theta_{ab}, Z_j^0) = \begin{cases} Z_j^0 + f_{ab}(t - E_j^\tau, \theta_{ab}, Z_j^0), & \text{If } I_j = 1, E_j = 1, \text{ and } t > E_j^\tau \\ Z_j^0, & \text{Otherwise} \end{cases} \quad (4)$$

Where $\theta_{ab} = \{a, b, c, \alpha\}$. Second, we define a likelihood function for the correlation of protection. For an individual j , with $E_j = 1$, the correlate of protection given exposure at time t with titre value $X_{j,t}$, given by a Bernoulli distribution with the probability is given by Equation 1. The PDF of this likelihood is given by Equation 5.

$$P_{cop}(I_j | Z_j^0, \theta_{cop}) = f_{cop}(Z_j^0, \theta_{cop})^{I_j} (1 - f_{cop}(Z_j^0, \theta_{cop}))^{1-I_j} \quad (5)$$

$\theta_{cop} = \{\beta_0, \beta_1\}$. Finally, we define an observational model to capture variability between hosts and measurement error. Given $X_{j,t}$ and the serological antibody data at the same time point is given by $Z_{j,t}$,

we assume the measurement error follows a normal distribution with a PDF given by Equation 6.

$$P_{obs}(Z_{j,t} | X_{j,t}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(Z_{j,t}-X_{j,t})^2}{2\sigma^2}\right)} \quad (6)$$

Let $\theta = \{a, b, c, \alpha, \beta_0, \beta_1, \sigma\}$ be the set of continuous parameters which are to be fitted in the model.

3.2 Posterior distribution via Bayes rule

We have two different likelihoods depending on whether an individual is exposed ($E_i = 1$) or not ($E_i = 0$).

3.2.1 Likelihood for an non exposed individual $E_j = 0$

In this case, the value of the timing of exposure and infection status is not applicable and thus not inferred. The likelihood for individual j with serological samples taken at times $t \in T_j$ is therefore equivalent to:

$$L_{E_j=0}(Z_j|\theta) = \prod_{t \in T_j} P_{obs}(Z_{j,t}|Z_j^0, \sigma) \quad (7)$$

as $X_{i,t} = Z_i^0$ for all t .

3.2.2 Likelihood for an exposed individual $E_j = 1$

In this case, the infectious status is determined by the correlation of the protection likelihood (P_{cop}) and the antibody kinetics. The likelihood for this individual with serological samples taken at times $t \in T_j$ and infection time E_j^τ is therefore equivalent to:

$$L_{E_j=1}(Z_j|I_j, \theta) = \prod_{t \in T_j} P_{obs}(Z_{j,t}|X_{j,t}, \sigma) P_{cop}(I_j | Z_j^0, \theta_{cop}) \quad (8)$$

where $X_{j,t} = F_{ab}(I_j, E_j^\tau, \theta_{ab}, Z_j^0)$.

3.2.3 Total likelihood

If \mathbf{E}_0 and \mathbf{E}_1 are vectors representing the set of individuals who are not exposed and exposed, respectively. Then, the total likelihood can be written

$$L(\mathbf{Z}|\mathbf{I}, \theta) = \prod_{j \in \mathbf{E}_0} L_{E_j=0}(Z_j|\theta) \prod_{j \in \mathbf{E}_1} L_{E_j=1}(Z_j|I_j, \theta) \quad (9)$$

3.2.4 Prior distributions

We choose prior distributions for each parameter $\pi(\theta)$. **Table 1** summarises the chosen priors with their support.

| Parameter | Prior (π) | Support (\mathcal{S}) |
|-----------|--------------------------|---------------------------|
| a | $\mathcal{N}(1.5, 0.5)$ | $[0.5, 4]$ |
| b | $\mathcal{N}(0.3, 0.05)$ | $[0, 1]$ |
| c | $\mathcal{U}(0, 4)$ | $[0, 4]$ |
| α | $\mathcal{U}(0, 1)$ | $[0, 1]$ |
| β_0 | $\mathcal{U}(-10, 10)$ | $[-10, 10]$ |
| β_1 | $\mathcal{U}(-10, 10)$ | $[-10, 10]$ |
| σ | $\mathcal{U}(0.01, 1)$ | $[0.01, 1]$ |

Table 1: Table with Headers: Parameter, Prior, and Support

We also choose the prior for the number of infections $n_{\mathbf{I}}$ given the number of exposed individuals $n_{\mathbf{E}}$ to be a Beta Binomial distribution: $\pi(\mathbf{I}) = \text{BetaBinomial}(n_{\mathbf{I}}|n_{\mathbf{E}}, 1, 1)$. Choosing this prior prevents any implicit priors that might arise from products of Bernoulli trials?? as $\text{BetaBinomial}(n_{\mathbf{I}}|n_{\mathbf{E}}, 1) = 1/n_{\mathbf{E}}$ for all $0 \leq n_{\mathbf{I}} \leq n_{\mathbf{E}}$.

3.2.5 Posterior distributions

Bayes' rule stipulates that the product of the prior distribution and likelihood is proportional to the posterior distribution; we can use this rule to approximate the posterior for use in the metropolis algorithm. Specifically

$$P(\theta, \mathbf{I}|\mathbf{Z}) \propto \mathcal{L}(\mathbf{Z}|\mathbf{I}, \theta)\pi(\theta)\pi(\mathbf{I}) \quad (10)$$

3.3 Metropolis-Hastings algorithm

3.3.1 Overview

The Metropolis-Hastings. (MH) algorithm is a widely used method for generating samples from a target probability distribution. It falls under the broader category of Markov Chain Monte Carlo (MCMC) methods and is particularly useful when direct sampling from the desired distribution is challenging or impossible such as the likelihood described above. The Metropolis-Hastings algorithm offers a solution to this problem. It is a Markov chain-based approach that iteratively generates a sequence of samples, which eventually converge to the desired distribution.

Say we wish to sample from an intractable probability distribution $P(x)$. The idea of the MH is to define a Markov chain so that the stationary distribution of the Markov chain is $P(x)$. That is, the resulting Markov chain from MH generates a sequence of values, denoted $\{x_1, x_2, \dots, x_n\}$, such that as $n \rightarrow \infty$ we can guarantee that $x_i \sim P(x)$. To do this, we uniquely define the Markov chain by its transition probabilities from x to x' , $F(x'|x)$, that must satisfy the detailed balance condition:

$$F(x' | x)P(x) = F(x | x')P(x') \quad (11)$$

This condition ensures that the i) probability density for the next step of the Markov chain is the same as the current density and that ii) this probability density is equal to the posterior. To construct a transition probability which satisfies this condition, we split P into a proposal distribution $q(x'|x)$ and an acceptance probability $\alpha(x, x')$:

$$F(x' | x) = q(x'|x)\alpha(x, x') \quad (12)$$

A common choice for $\alpha(x, x')$ which satisfies the detailed balance condition, is the acceptance ratio given by

$$\alpha(x, x') = \min \left(1, \frac{P(x')}{P(x)} \cdot \frac{Q(x | x')}{Q(x' | x)} \right) \quad (13)$$

With this, the user has a choice over the proposal distribution Q , which can be tailored to optimise the general algorithm given in **Algorithm 1**.

Algorithm 1 Generic Metropolis-Hastings Algorithm

```

1: Initialize the chain with an initial state  $\theta^{(0)}$ 
2: for  $i = 1$  to  $N$  do
3:   Generate a candidate state  $\theta'$  from the proposal distribution:  $\theta' \sim Q(\cdot | \theta^{(i)})$ 
4:   Compute the acceptance ratio:


$$\alpha(\theta^{(i)}, \theta') = \min \left( 1, \frac{P(\theta')}{P(\theta^{(i)})} \cdot \frac{Q(\theta^{(i)} | \theta')}{Q(\theta' | \theta^{(i)})} \right)$$


5:   Sample  $u \sim \mathcal{U}(0, 1)$ 
6:   if  $u \leq \alpha$  then
7:     Accept the candidate state:  $\theta^{(i+1)} \leftarrow \theta'$ 
8:   else
9:     Reject the candidate state:  $\theta^{(i+1)} \leftarrow \theta^{(i)}$ 
10:  end if
11: end for

```

3.3.2 MH for serological inference with known exposure

In our model, we wish to sample from the posterior density function given by **Equation 10**, which infers θ , and infectious statuses $I_j \in \mathbf{I}$, for $1 \leq j \leq M$ individuals. For the proposal distribution, we define independent proposal distribution for θ and \mathbf{I} , such that $Q(\theta, \mathbf{I}) = q_\theta(\theta)q_I(\mathbf{I})$. At Markov chain step i , we have a value of the parameter space, $\theta^{(i)}$, and propose a new set of parameters θ' via the proposal distribution $\theta \sim q_\theta(\cdot | \theta^{(i)}, \psi_{adapt}^{(i)})$. This proposal is a multivariate normal distribution with an adaptive covariance matrix, which is defined by the set of parameters $\psi_{adapt}^{(i)}$, which are updated at each time step.[ref] (See Appendix). For \mathbf{I} , we propose a new infection state \mathbf{I}' by selecting an exposed individual j , which has infection status $I_j^{(i)}$ at step i of the current Markov chain, we sample a proposal value for their infection status I'_j by the proposal distribution for $I'_j \sim q_I(\cdot | I_j^{(i)}) = \text{Bernoulli}(0.5)$. Therefore the proposal for $q_I(\mathbf{I}' | \mathbf{I}) = 1/n_E 0.5$ for all j . Both of these proposals $q_\theta(\theta | \theta^{(i)}, \psi_{adapt}^{(i)})$, $q_I(\mathbf{I}' | \mathbf{I})$ are both symmetric and thus cancel out the acceptance ratio (**Equation 13**). Further, the prior distribution $\pi(\mathbf{I}) = 1/n_E$ for all $0 \leq n_I \leq n_E$, and thus also cancels out in the acceptance ratio, therefore we need only calculate: $P(\theta, \mathbf{I} | \mathbf{Z}) \propto \mathcal{L}(\mathbf{Z} | \mathbf{I}, \theta)\pi(\theta)$

Consequently, we construct a new algorithm for inference of the known exposure model (**Algorithm 2**).

Algorithm 2 Metropolis-Hastings Algorithm for antibody kinetics and infection inference

```

1: Initialize the chain with an initial state  $\theta^{(0)}$  from the priors  $\pi(\cdot)$  and  $I_j^{(0)} \sim \text{Bernoulli}(0.5)$  for all  $1 \leq j \leq M$ 
   individuals to initialise  $\mathbf{I}^{(0)}$ , and initialise  $\psi_{adapt}^{(0)}$ .
2: for  $i = 1$  to  $N$  do
3:   Generate a candidate state  $\theta' \sim q_\theta(\theta^{(i)}, \psi_{adapt}^{(0)})$ 
4:   Generate a candidate individual  $j \in \mathbf{E}_1$ , then a candidate state  $I'_j \sim \text{Bernoulli}(0.5)$  to propose  $\mathbf{I}'$ 
5:   Compute the acceptance ratio:

$$\alpha((\theta^{(i)}, \mathbf{I}^{(i)}), (\theta', \mathbf{I}')) = \min \left( 1, \frac{P(\theta', \mathbf{I}' | Z)}{P(\theta^{(i)}, \mathbf{I}^{(i)} | Z)} \right)$$

6:   Sample  $u \sim \mathcal{U}(0, 1)$ 
7:   if  $u \leq \alpha$  then
8:     Accept the candidate state:  $\theta^{(i+1)} \leftarrow \theta'$  and  $\mathbf{I}^{(i+1)} \leftarrow \mathbf{I}'$ 
9:   else
10:    Reject the candidate state:  $\theta^{(i+1)} \leftarrow \theta^{(i)}$  and  $\mathbf{I}^{(i+1)} \leftarrow \mathbf{I}^{(i)}$ 
11:   end if
12:   Update  $\psi_{adapt}^{(i+1)} \leftarrow \psi_{adapt}^{(i)}$ 
13: end for

```

3.4 Implementation

Algorithm 2 is coded manually in R and Rcpp. We run the algorithm for four chains, each with 200,000 steps, with 100,000 burn-ins steps. The initial values for θ and \mathbf{I} are their prior distributions. We initialise the adaptive covariance by running with an identity matrix with each parameter scale according to 1,000 steps, then sample from the adaptive scheme as in XX. (Appendix). We thin the posterior samples by taking one in every 100 samples, leaving 1,000 posterior samples.

3.5 Simulation recovery

After running **Algorithm 2**, we plot the posterior samples, $\hat{\theta}$ and $\hat{\mathbf{I}}$ and compare with the simulated parameters.

3.5.1 Infection recovery

We assess the ability of the algorithm to recover the infection status of each individual in the study. If the set posterior samples of the infection status for individual j is given by \hat{I}_j , then we plot the expectation $\mathbb{E}(\hat{I}_j)$ so we can assess the ability of the algorithm to recover the individual-level simulated infection status (**Figure 9a**). Given no COP model A, we find when the pre-infection titre < 3.3 log titre value that all six models considered can recover the infection status of almost all individuals. When the pre-infection titre is greater than 3.3, the attenuation of boosting for infected individuals causes no meaningful change in the antibody kinetics ($f_{ab}^2(Z, \alpha) = 0$ when $Z > 3.3$). Thus, these individuals' infections are difficult to infer serologically as their titre dynamics are equivalent to independent of their infection status. In our COP model B, we find that including the correlation of protection influences the infection status. As the inferred correlate has a low probability of infection at higher titres, this causes the \hat{I}_j to be more likely to be 0 at higher titre values. Thus, the infection statuses for nearly all individuals are recoverable for COP model B.

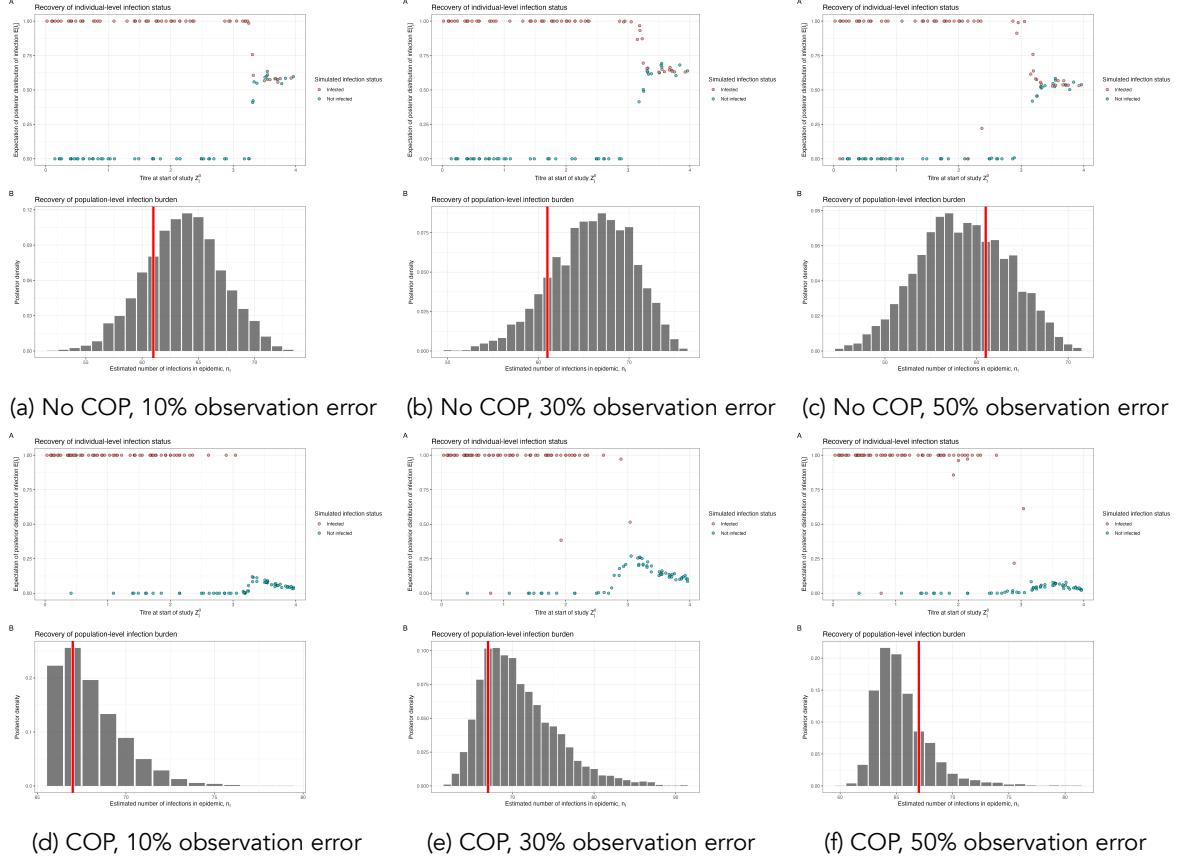


Figure 4: Simulation recovery of the individual infection status, \hat{I}_j , for two COP models (top: No COP, bottom: logistic COP) and three different levels antibody kinetics variability (10%, 30%, 50%)

3.5.2 Correlate of protection

We next assess the ability of **Algorithm 2** to recover the correlate of protection function $f_{cop}(x, \hat{\theta}_{cop})$, where x is the titre value at infection and where $\hat{\theta}_{cop} = \{\hat{\beta}_0, \hat{\beta}_1\}$ are the posterior samples for β_0 and β_1 . We consider two COP models: COP model A, no correlate of protection, and COP model B, a logistic curve for COP. For Model A, we find that the COP curve is mostly recovered, with the simulated line within a 50% confidence interval of the posterior sample (**Figure ??**). For Model B, we find the logistic shape of the COP is recovered in the posterior samples. The variability in the antibody kinetics seemed to have a negligible effect on the recoverability of the COP curve. To understand the difference between the simulated functional form in **Figure ??** in red, and the posterior samples, we have plotted the inferred COP from the simulated infection states in **Figure 3**. Here, it is clear, although we have a pre-defined function for the correlate of protection, when we simulate the data, the COP curve is not perfectly recovered, and thus, the posterior distribution of our inference method can, at best, recover the inferred COP curve.

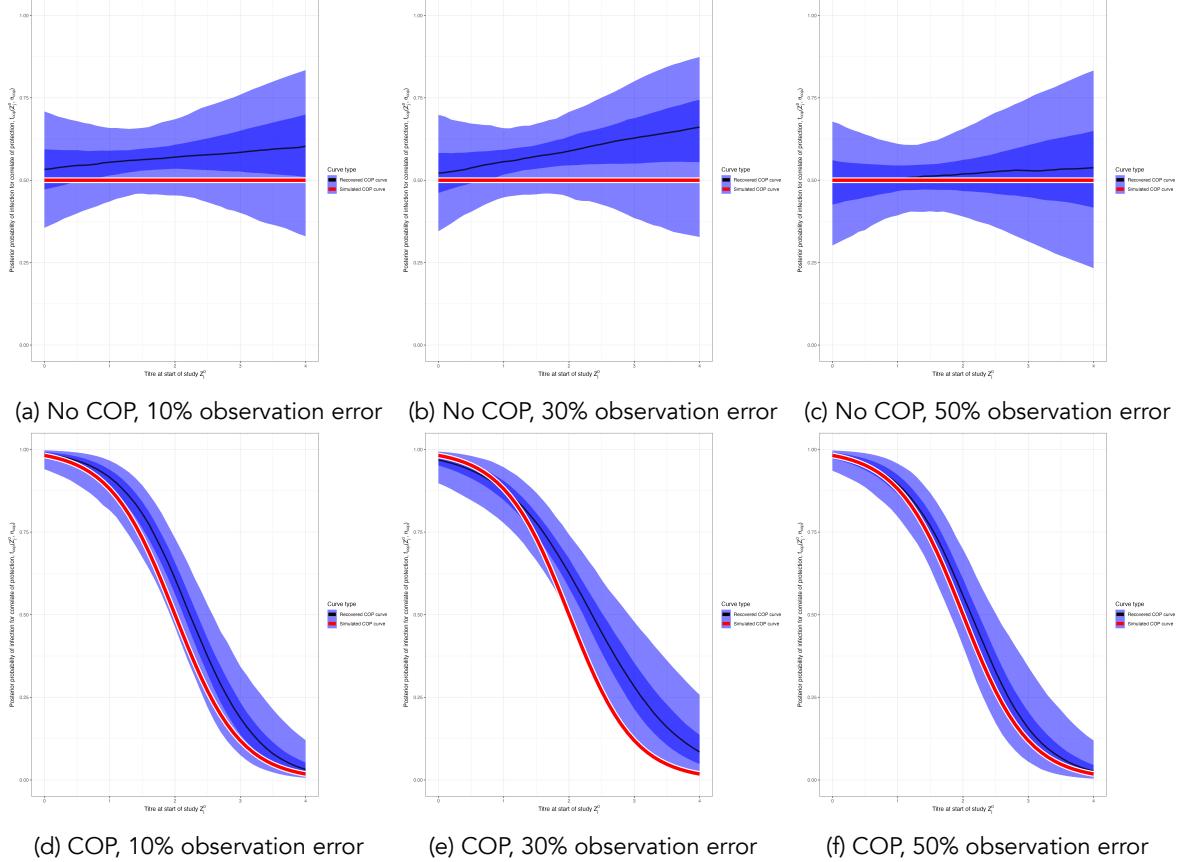


Figure 5: Simulation recovery of the COP function, with posterior samples plot $f_{cop}(x, \hat{\theta}_{cop})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).

3.5.3 Antibody kinetics

Algorithm 2 also successfully recovers the simulated antibody kinetics. Let us plot $f_{ab}^1(s, \hat{a}, \hat{b}, \hat{c})$, the posterior predictive distribution for the antibody kinetic boosting, given posterior distributions for \hat{a} , \hat{b} , and \hat{c} . At all three levels of kinetic uncertainty, the antibody kinetics are recovered, though increasing uncertainty weakens the accuracy of the recovered curves compared to the simulated. (**Figure ??**).

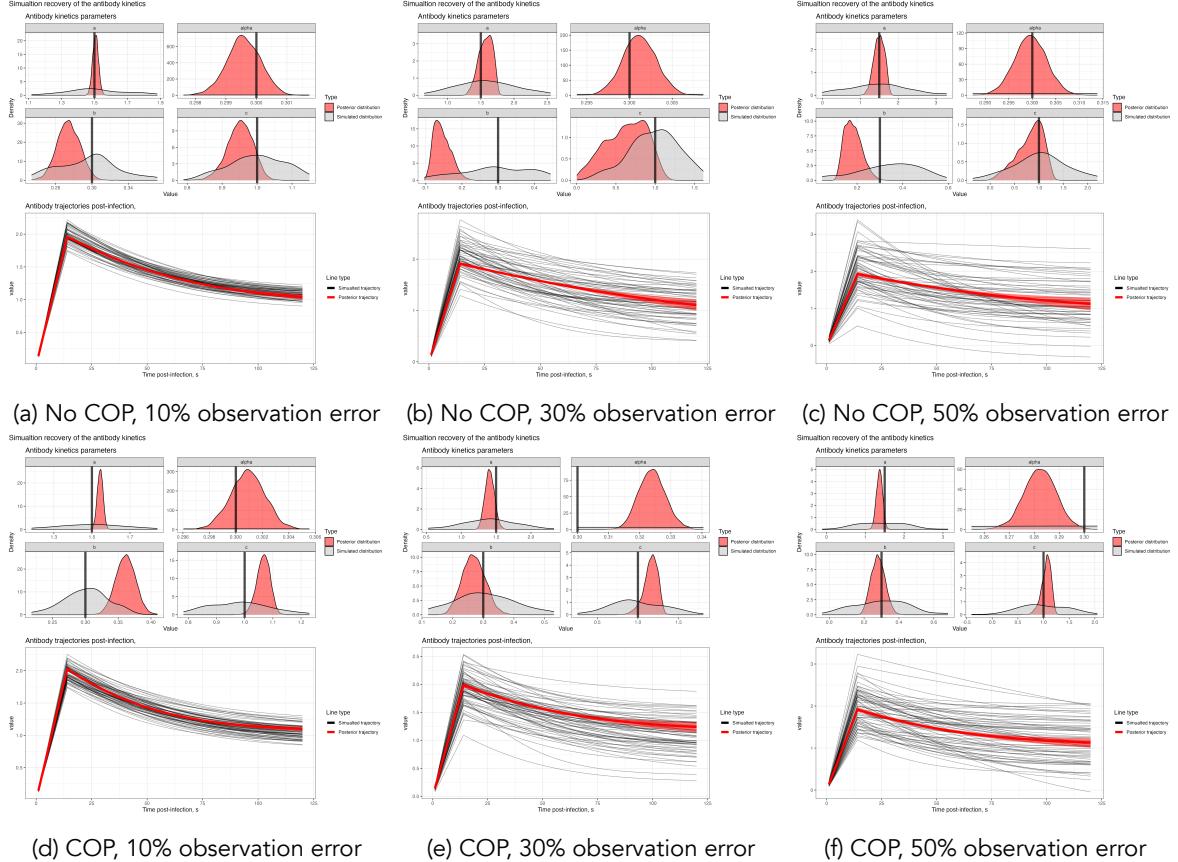


Figure 6: Simulation recovery of the antibody kinetics function with posterior samples plot $f_{ab}^1(s, \hat{a}, \hat{b}, \hat{c})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).

4 Inference with unknown exposure status

For the known exposure status of an individual, we have shown that **Algorithm 2** can recover the individual-level infection status, a population-level correlate of protection and the underlying antibody kinetics for two different correlate of protection assumptions and three different levels of individual-level kinetics variability. In practice, this algorithm is unlikely to be useful as the individual-level exposure state is unknown. In this section, we will expand on this algorithm for the case when exposure status is unknown throughout the serosurvey.

4.1 Overview

In the case where the exposure status of each individual, j , is unknown, we must now infer their exposure state $E_j \in \{0, 1\}$ and the time of exposure given they are exposed $0 \leq E_j^\tau \leq 120$. In the case where $E_j = 0$, the likelihood is as derived in **Equation ??**. However, in the case where $E_j = 1$, the likelihood contains an additional dependencies:

$$L_{E_j=1}(Z_j|I_j, E_j^\tau, \theta) = \prod_{t \in T_j} P_{obs}(Z_{j,t}|X_{j,t}, \sigma)P_{cop}(I_j | E_j^\tau, \theta) \quad (14)$$

where $X_{j,t} = P_{ab}(I_j, E_j^\tau, \theta_{ab}, Z_i^0)$.

4.1.1 Likelihood and priors

Let $\mathbf{E} = \{E_0, E_1, \dots, E_M\}$ be a vector describing the exposure status of each individual, let $\mathbf{E}^\tau = \{E_0^\tau, E_1^\tau, \dots, E_{n_{\mathbf{E}_1}}^\tau\}$ be a vector describing the exposure times for each exposed individual. The likelihood of this system is similar to before:

$$\mathcal{L}(\mathbf{Z}|\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}) = \prod_{j \in \mathbf{E}_0} L_{E_j=0}(Z_j|\theta) \prod_{j \in \mathbf{E}_1} L_{E_j=1}(Z_j|I_j, E_j^\tau, \theta) \quad (15)$$

We now must additionally define priors for $\pi(\mathbf{E})$ and $\pi(\mathbf{E}^\tau)$. Similar to $\pi(\mathbf{I})$ we define the prior distribution for $\pi(\mathbf{E})$ to be a Beta Binomial distribution: $\pi(\mathbf{E}) = \text{BetaBinomial}(n_{\mathbf{E}}|M, 1, 1)$, which is equal to $1/M$ for all $0 \leq n_{\mathbf{E}} \leq M$. For $\pi(\mathbf{E}^\tau)$, we assume that each element $E_j^\tau \in \mathbf{E}^\tau$ has a prior given by P_t such that $\pi(\mathbf{E}^\tau) = \prod_{j=1}^{n_{\mathbf{E}_1}} P_t(E_j^\tau)$. The priors for $\pi(\theta)$ and $\pi(\mathbf{I})$ are as described in **Section 2**.

Consequently, we sample from the posterior distribution

$$P(\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}|\mathbf{Z}) \propto \mathcal{L}(\mathbf{Z}|\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I})\pi(\theta)\pi(\mathbf{E})\pi(\mathbf{E}^\tau)\pi(\mathbf{I}) \quad (16)$$

If we use a **Algorithm 2** or any Metropolis Hasting algorithm to infer the exposure status and exposure time, we run into a problem. The number of parameters in the posterior distribution changes according to whether an individual is exposed, as those who are exposed, have parameters I_j and E_j^τ to infer, whereas an individual who is not exposed has neither. Therefore, regardless of the proposal distribution we choose for inferring \mathcal{E} , we cannot use the existing algorithm highlighted in **Algorithm 2** as the detailed balance condition now fails.

4.2 The Reversible-Jump MCMC

The Reversible Jump Markov Chain Monte Carlo (MCMC) algorithm is a Bayesian statistical method designed for model selection in situations where the number of model parameters can vary. It achieves this by introducing a stochastic mechanism that proposes moves between different models, including adding or removing parameters. The idea is to use a Metropolis-Hastings step to evaluate the acceptance probability of these proposed model changes, ensuring that the Markov chain explores the posterior distribution over both model parameters and model structures.

4.2.1 Mathematical overview

Let $\{k \in \mathcal{K}\}$ denote a collection models and θ_k be the parameter space of model k . A full Bayesian model for inferring k and θ_k can be written:

$$p(k)p(\theta_k|k)p(Z|k, \theta_k)$$

where $p(k)$ is the prior probability that model k is chosen, $p(\theta_k|k)$ is the prior distribution for parameters θ_k in model k , and $p(Z|k, \theta_k)$ is the likelihood for the observed data for model k . We wish to build a Markov chain Monte Carlo algorithm to sample from the stationary distribution:

$$P(\theta_k, k|Z) \propto p(k)p(\theta_k|k)p(Z|\theta_k, k) \quad (17)$$

However, as the dimensions of vector θ_k is changing as we switch between models with different dimensions, there is no way obvious way to define Q and α such that the detailed balance condition (**Equation 11**) is met. That is, the posterior density for proposal state cannot be the same as the current density as the dimensions have changed. Therefore, the sampler is not converging to a single posterior distribution.

The reversible jump mcmc proposal a solution to this issue **??**. The idea is to augment both the current state and the proposed state with sampled parameters, then define a bijection between these two augmented spaces, and then redefine α such that the detailed balance condition holds. Let $x = (k, \theta_k)$ denote the model number k and θ_k the parameters associated with model k ($\theta_k \in \mathbb{R}^{d_k}$) then define the proposed state as $x' = (k', \theta_{k'})$, with $\theta_{k'} \in \mathbb{R}^{d_{k'}}$). We write the proposal $Q(x'|x)$, the probability of moving to state x' from state x in the form

$$Q(x'|x) = Q((k', \theta_{k'})|(k, \theta_k)) = q_X(\theta_{k'}|\theta_k, k', k) \cdot q_k(k'|k) \quad (18)$$

where $q_k(k'|k)$ is the probability of selecting model k' from model k and q_X the probabiltiy of sampling $\theta_{k'}$ given current parameters θ_k and known k , and known proposed model k' . The challenge with q_X is that we must adjust for the change in dimensions of the parameter space of $\theta_{k'}$ compared to θ_k (i.e $d_k \neq d_{k'}$). To do this, we sample auxiliary variables to match the dimensions and define a bijection between the augmented spaces. Thus if $d_k \neq d_{k'}$, we generate a random variables of length s , $\mathbf{u} = (u_1, \dots, u_s) \sim q_1(\mathbf{u})$ and one of length s' , $\mathbf{u}' = (u'_1, \dots, u'_{s'}) \sim q_2(\mathbf{u}')$ such that $d_{k'} + s' = d_k + s$. We then define a bijection, T

$$(\theta_{k'}, \mathbf{u}') = T(\theta_k, \mathbf{u}) \quad (19)$$

to ensure the reversibility of the proposal distribution.

For the detailed balance condition to hold, Green**??** shows a prospal distribution given by

$$Q(x|x') = q_k(k|k')q_X(\theta_{k'}|\theta_k, k, k') = q_k(k|k')q_2(\mathbf{u}') \left| \frac{\partial(\theta_{k'}, \mathbf{u}')}{\partial(\theta_k, \mathbf{u})} \right|$$

$$Q(x'|x) = q_k(k'|k)q_X(\theta_k|\theta_{k'}, k, k') = q_k(k'|k)q_1(\mathbf{u})$$

where $\left| \frac{\partial(\theta_{k'}, \mathbf{u}')}{\partial(\theta_k, \mathbf{u})} \right|$ is the jacobian of the transformation T . Then, choosing an acceptance ratio given

$$\alpha(x, x') = \min \left(1, \frac{P(x)q_k(k|k')q_2(\mathbf{u}')}{P(x')q_k(k'|k)q_1(\mathbf{u})} \cdot \left| \frac{\partial(\theta_{k'}, \mathbf{u}')}{\partial(\theta_k, \mathbf{u})} \right| \right) \quad (20)$$

ensures the stationary distribution chain samples

$$P(\theta_k, k|Z) \propto p(k)p(\theta_k|k)p(Z|\theta_k, k) \quad (21)$$

A general form of the RJMCMC then follows Algorithm 3.

Algorithm 3 Reversible Jump MCMC Algorithm

```

1: Chose a model  $k$ 
2: Initialize the chain with an initial state  $\theta_k^{(0)}$ 
3: for  $i = 1$  to  $N$  do
4:   Sample model  $k' \sim q(\cdot|k^{(i)})$ 
5:   Sample  $\mathbf{u} \sim q_2(\mathbf{u})$ 
6:   Set  $(\theta_{k'}, \mathbf{u}') = T(\theta_k^{(i)}, \mathbf{u})$ 
7:   Compute the acceptance ratio:
```

$$\alpha((k^{(i)}, \theta_k^{(i)}), (k', \theta_{k'})) = \min \left(1, \frac{P(k', \theta_{k'} | Z) q(k^{(i)} | k') q_2(\mathbf{u}')}{P(k^{(i)}, \theta_k^{(i)} | Z) q(k' | k^{(i)}) q_1(\mathbf{u})} \cdot \left| \frac{\partial(\theta_{k'}, \mathbf{u}')}{\partial(\theta_k^{(i)}, \mathbf{u})} \right| \right)$$

```

8:   Generate a uniform random number  $u$  from the interval  $[0, 1]$ 
9:   if  $u \leq \alpha$  then
10:    Accept the candidate state:  $k^{(i+1)} \leftarrow k'$  and  $\theta^{(i+1)} \leftarrow \theta_{k'}$ 
11:   else
12:    Reject the candidate state:  $k^{(i+1)} \leftarrow k^{(i)}$  and  $\theta^{(i+1)} \leftarrow \theta^{(i)}$ 
13:   end if
14: end for
```

4.3 Application of RJMCMC to serological data

Algorithm 3 is a general framework for jumping from a model k with a parameter values $\theta_k \in \Theta_k$ and another model k' with a parameter values $\theta_{k'} \in \Theta_{k'}$. For our serological inference, our model k , represent different elements of the exposure state vector $\mathbf{E}_k = \{E_0, E_1, \dots, E_{n_{\mathbf{E}_k}}\}$ (let $|\mathbf{E}_k| = n_{\mathbf{E}_k}$ be the number of exposed individuals and $n_{\mathbf{E}_0} = M - n_{\mathbf{E}_1}$ be the number of non-exposed individuals in model k). For a given exposure vector \mathbf{E}_k , we define three different possible ways to sample a new exposure state vector, \mathbf{E}' in our RJMCMC algorithm: a birth move (adding a new exposure), death move (remove an existing exposure), and parameter updating (exposure state remains the same)[REF].

4.3.1 Birth move

For a given exposure vector \mathbf{E}_k , a birth move generates a new exposure vector \mathbf{E}' , by randomly selecting a non-exposed individual and changing their exposure status from $E_j = 0 \rightarrow E'_j = 1$ and appending it to \mathbf{E}_k . We can derive an expression for $q_k(k'|k)$ by separating into the probability that a birth move is selected at model k , $q_{birth}(k'|k)$ and the probability of choosing individual j uniformly from the non-exposed individuals:

$$q_k(k'|k) = q_{birth}(k'|k) \cdot \frac{1}{n_{\mathbf{E}_0}} \quad (22)$$

To understand how to evaluate $q_2(\mathbf{u}')$, $q_1(\mathbf{u})$, consider the change in likelihood for an individual j who is chosen to be exposed:

$$\prod_{t \in T} P_{obs}(Z_{j,t} | Z_j^0, \sigma) \rightarrow \prod_{t \in T} P_{obs}(Z_{j,t} | X_{j,t}, \sigma) P_{cop}(I_j | Z_j^0, \theta) P_t(E_j^t) \quad (23)$$

with the likelihood function staying the same for all other individuals. By changing the exposure state for individual j , the likelihood now depends on two parameters not in the previous likelihood: the timing of the exposure E_j^t and their infection status I_j . In the notation of the previous section, it is convenient to define $\mathbf{u} = (E_j^t, I_j)$. Therefore, we must define a sampling procedure for \mathbf{u} and a probability density function $q_1(\mathbf{u})$ (Note as $d_{k'} > d_k$, we can assume \mathbf{u}' is empty). For E_j^t , we sample from the probability density function for the timing $P_t(\cdot)$, and for the infection status, we sample from $P_{cop}(\cdot | Z_j^0, \theta_{cop})$ (see **Equation 5**).

This sampling procedure results in a proposed sample which is in the proposed parameter space $\Theta_{k'} : (\theta_k, E_j^\tau, I_j) = \theta_{k'} \in \Theta_{k'}$. Consequently, we can choose the identity function for the required bijection T in **Equation 19**, which means the Jacobian is equal to 1.

With this sampling proposal, we can then evaluate $q_1(\mathbf{u})$ through likelihood functions for each of I_j and E_j^τ :

$$q_1(\mathbf{u}) = q_1(E_j^\tau, I_j) = P_{cop}(I_j | E_j^\tau, \theta_{cop}) P_t(E_j^\tau) \quad (24)$$

Now let us consider the reverse move, which is from the proposed model k' moving back to the model k . For this move, we randomly select one person from the proposed exposure state \mathbf{E}' and change their exposure state from $E_j = 1 \rightarrow E'_j = 0$. Similar to above, we derived an expression for $q_k(k|k')$ by separating into the probability that an inverse birth move is selected at model k , $q_{birth}(k|k')$ and the probability of choosing individual j uniformly from the exposed individuals of \mathbf{E}' :

$$q_k(k|k') = q_{birth}(k|k') \cdot \frac{1}{1 + n_{\mathbf{E}_1}} \quad (25)$$

In this move, we are removing the parameters E_j^τ and I_j from $\theta_{k'}$. After this, we are left with $\theta_k \in \Theta_k$, so we do not need to sample new variables to generate samples from Θ_k . Thus \mathbf{u}' is empty and $q_2(\mathbf{u}') = 1$. The acceptance ratio (**Equation 20**) for a birth move, where our current state is $k = \{\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}\} \rightarrow k' = \{\theta, \mathbf{E}', \mathbf{E}'^\tau, \mathbf{I}'\}$ is updated according to a uniformed sampled non-exposed individual j :

$$\alpha(k, k') = \min \left(\frac{P(\theta, \mathbf{E}', \mathbf{E}'^\tau, \mathbf{I}', |Z) q_{birth}(k|k') n_{\mathbf{E}_0}}{P(\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}, |Z) P_{cop}(I'_j | E'_j, \theta_{cop}) P_t(E'_j) q_{birth}(k'|k) (n_{\mathbf{E}_1} + 1)} \right) \quad (26)$$

4.3.2 Death move

For a given exposure vector \mathbf{E}_k , a death move generates a new exposure vector \mathbf{E}' , by randomly selecting an exposed individual and changing their exposure status from $E_j = 1 \rightarrow E'_j = 0$ therefore removing it from \mathbf{E}_k . We can derive an expression for $q_k(k'|k)$ by separating into the probability that a death move is selected at model k , $q_{death}(k'|k)$ and the probability of choosing individual j uniformly from the exposed individuals:

$$q_k(k'|k) = q_{death}(k'|k) \cdot \frac{1}{n_{\mathbf{E}_1}} \quad (27)$$

The reverse probability $q_k(k'|k)$ is equivalent to a 'birth' move, that is the probability of sampling a non-exposed person in \mathbf{E}' , or

$$q_k(k|k') = q_{death}(k|k') \cdot \frac{1}{1 + n_{\mathbf{E}_0}} \quad (28)$$

To understand how to evaluate $q_1(\mathbf{u}), q_2(\mathbf{u}')$, consider the change in likelihood for an individual j who is chosen to be exposed:

$$\prod_{t \in T} P_{obs}(Z_{j,t} | X_{j,t}, \sigma) P_{cop}(I_j | Z_j^0, \theta) P_t(E_j^\tau) \rightarrow \prod_{t \in T} P_{obs}(Z_{j,t} | Z_j^0, \sigma) \quad (29)$$

with the likelihood function staying the same for all other individuals. By changing the exposure state for individual j , the likelihood now depends on two fewer parameters than were in the previous likelihood for k' : the timing of the exposure E_j^τ and their infection status I_j . Therefore, by using the same argument as in the 'birth move' section, defining $q_1(\mathbf{u}) = 1$ and $q_2(\mathbf{u}') = F_t(E_j^\tau) F_{cop}(I_j | E_j^\tau, \theta_{cop})$, we can take the identity bijection as sample a value in state $\theta' \in \Theta'$ directly. The acceptance ratio (**Equation 20**) for a death

move, where our current state is $k = \{\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}\} \rightarrow k' = \{\theta, \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}'\}$ is updated according to a uniformed sampled exposed individual j :

$$\alpha(k, k') = \min \left(\frac{P(\theta, \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}, |Z) P_{cop}(I_j | E_j, \theta_{cop}) P_t(E_j^\tau) q_{death}(k|k') n_{\mathbf{E}_1}}{P(\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}, |Z) q_{death}(k'|k) (n_{\mathbf{E}_0} + 1)} \right) \quad (30)$$

4.3.3 Parameter updating

In this case, $\mathbf{E}_1 = \mathbf{E}'$, that is the exposure vector remains unchanged, with probability $q_{par}(k|k)$. Therefore, the detailed balanced conditions are met without dimensional adjustment. In this case, we can use sample values for θ and \mathcal{I} and use the acceptance ratio highlighted in Algorithm 3.

Note on q_{birth} , q_{death} , q_{par} : We can select values $q_{birth}(k'|k)$ and $q_{death}(k'|k)$ which simplify the expression in the acceptance ratios in **Equation 26** and **Equation 30**. If k is such that $n_{\mathbf{E}} = 0$, then select we choose $q_{par} = 1/3$ and $q_{birth} = 2/3$. If $n_{\mathbf{E}} = M$, then we choose $q_{par} = 1/3$ and $q_{death} = 2/3$. Otherwise, we $q_{birth}(k'|k) = q_{death}(k'|k) = q_{par} = 1/3$ for all k' . Choosing these values means that the values of q_{birth} , q_{death} , and q_{par} cancel in all acceptance ratios (**Equation 26** and **Equation 30**) for all values of k' given k .

An algorithm describing the Birth-Death RJMCMC algorithm for this data is given in **Algorithm 4**.

Algorithm 4 Birth-Death Reversible Jump MCMC Algorithm

1: Choose a model k and initialize the chain with an initial states $\theta_k^{(0)}$, $\mathbf{E}^{(0)}$, $\mathbf{E}^{\tau,(0)}$ and $\mathbf{I}^{(0)}$. If $0 < n_{\mathbf{E}} < M$, then $p_{birth} = p_{death} = p_{par} = 0.33$; if $n_{\mathbf{E}} = 0$, $p_{birth} = 0.67, p_{par} = 0.33, p_{death} = 0$; if $n_{\mathbf{E}} = M$, $p_{death} = 0.67, p_{par} = 0.33, p_{birth} = 0$.
 2: **for** $i = 1$ to N **do**
 3: $u_1 \sim \mathcal{U}(0, 1)$
 4: **if** $u_1 \leq p_{birth}$ **then**
 5: Birth move. Select $j'' \in \mathbf{E}_0^{(i)}$, set $E_{j''} = 1$, sample $E_{j''}^\tau \sim P_t(\cdot)$, $I_{j''} \sim P_{cop}(\cdot | Z_{j''}^0, \theta_{cop})$ and update $\{\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)}\} \rightarrow \{\theta', \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}'\}$. Then calculate the acceptance probability

$$\alpha(k^{(i)}, k') = \min \left(\frac{P(\theta', \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}', | Z) n_{\mathbf{E}_0}}{P(\theta, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)}, | Z) P_{cop}(I_{j''} | E_{j''}, \theta_{cop}) P_t(E_{j''}^\tau) (n_{\mathbf{E}_1} + 1)} \right)$$
 6: **else if** $u_1 \leq (p_{birth} + p_{death})$ **then**
 7: Death move. Select $j' \in \mathbf{E}_1^{(i)}$, set $E_{j'} = 0$ and update $\{\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)}\} \rightarrow \{\theta', \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}'\}$. Then calculate the acceptance probability

$$\alpha(k^{(i)}, k') = \min \left(\frac{P(\theta, \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}' | Z) P_{cop}(I_{j'} | E_{j'}, \theta_{cop}) P_t(E_{j'}^\tau) n_{\mathbf{E}_1}}{P(\theta, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)}, | Z) (n_{\mathbf{E}_0} + 1)} \right)$$
 8: **else**
 9: Sample a candidate state $\theta' \sim q_\theta(\theta^{(i)}, \psi_{adapt}^{(i)})$
 10: Sample $j' \in \mathbf{E}_1^{(i)}$, and then a candidate state $I_j' \sim \text{Bernoulli}(0.5)$
 11: Compute the acceptance ratio:

$$\alpha(k^{(i)}, k') = \min \left(1, \frac{P(\theta', \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}' | Z)}{P(\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)} | Z)} \right)$$
 12: Update $\psi_{adapt}^{(i+1)} \leftarrow \psi_{adapt}^{(i)}$
 13: **end if**
 14: Sample $u \sim \mathcal{U}(0, 1)$
 15: **if** $u \leq \alpha$ **then**
 16: Accept the candidate state: Let $\{\theta^{(i+1)}, \mathbf{E}^{(i+1)}, \mathbf{E}^{\tau,(i+1)}, \mathbf{I}^{(i+1)}\} \leftarrow \{\theta', \mathbf{E}', \mathbf{E}^{\tau'}, \mathbf{I}'\}$
 17: **else**
 18: Reject the candidate state: Let $\{\theta^{(i+1)}, \mathbf{E}^{(i+1)}, \mathbf{E}^{\tau,(i+1)}, \mathbf{I}^{(i+1)}\} \leftarrow \{\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)}\}$
 19: **end if**
 20: **end for**

Note on prior distributions As M is fixed, then $\pi(\mathbf{E}) = 1/M$ for all $0 \leq n_{\mathbf{E}} \leq M$ and thus cancels out in the acceptance ratio for the birth, death and parameter update move. $\pi(\mathbf{I})$ cancels out in the parameter updating acceptance ratio (as described in **Algorithm 2**). However, in the birth and death move, as $n_{\mathbf{E}}$ in the current state and $n_{\mathbf{E}'}$ is the proposed state have different values, then $\pi(\mathbf{I}) = 1/n_{\mathbf{E}_1} \neq 1/n_{\mathbf{E}_1'} = \pi(\mathbf{I}')$ no longer cancels in the ratio and must be included to ensure the detailed balance condition holds.

4.3.4 Within Gibbs sampling of exposure times

Algorihtm 4 allows for efficient sampling of the θ , \mathbf{I} , and \mathbf{E} . However, the timing of the exposures, \mathbf{E}^τ is not efficiently explored as it can only be changed at a birth or death move. Thereforer, it is desirable that values of \mathbf{E}^τ can be explored for fixed values of \mathbf{I} , and \mathbf{E} . To do this, we modify **Algorihtm 4** to allow the possibility of exploration of the \mathbf{E}^τ timings for a given model k whilst \mathbf{I} , and \mathbf{E} remain fixed. To implement this, after the candidate state has been accepted or rejected in **Algorithm 4**, we then resample a proportion of the \mathbf{E}^{tau} , and for each individual, we sample a new time E_j^τ from the proposal distribution

$$E_{j'}^\tau \sim q_t(E_j^\tau)$$

where we choose the proposal to be the symmetric $q_t(E_j^\tau) = \text{mathcalN}(E_j^{t,(i)}, \sigma_j^{(i)})$. Where $\sigma_j^{(i)}$ is an adaptively updated standard deviation for the proposal for the normal, which updates according to the regime:

$$\log(\sigma_j^{(i+1)}) = \log(\sigma_j^{(i)}) + (1+i)^{-0.5} * (\alpha - 0.44)$$

where α is the metropolis hasting ratio for $E_{j'}^\tau$ vs $E_j^{\tau,(i)}$.

Therefore the final RJMCMC algorithm (**Algorithm 5**) which effectively samples values from $\theta, \mathbf{E}, \mathbf{E}_\tau, \mathbf{I}$.

Algorithm 5 Efficient Birth-Death Reversible Jump MCMC Algorithm

```

1: Chose a model  $k$  and initialize the chain with an initial states  $\theta_k^{(0)}, \mathbf{E}^{(0)}, \mathbf{E}^{\tau,(0)}$  and  $\mathbf{I}^{(0)}$ .
2: for  $i = 1$  to  $N$  do
3:   Update  $\{\theta^{(i+1)}, \mathbf{E}^{(i+1)}, \mathbf{E}^{\tau,(i+1)}, \mathbf{I}^{(i+1)}\}$  according to Algoithm 4.
4:   for  $k = 1$  to  $N_k$  do
5:     Select  $j' \in N_{E=1}$  and resample from the proposal  $E_{j'}^\tau \sim \mathcal{N}(E_j^{\tau,(i)}, \sigma_j^{(i)})$  and update  $\mathbf{E}^{\tau,(i)} \rightarrow \mathbf{E}^{\tau,*}$ 
6:     Compute the acceptance ratio:


$$\alpha(k^{(i)}, k') = \min \left( 1, \frac{P(\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau'}, \mathbf{I}^{(i)} | Z)}{P(\theta^{(i)}, \mathbf{E}^{(i)}, \mathbf{E}^{\tau,(i)}, \mathbf{I}^{(i)} | Z)} \right)$$


7:     Sample  $u \sim \mathcal{U}(0, 1)$ 
8:     if  $u \leq \alpha$  then
9:       Accept the candidate state:  $\mathbf{E}^{\tau,(i+1)} \leftarrow \mathbf{E}^{\tau'}$ 
10:      else
11:        Reject the candidate state:  $\mathbf{E}^{\tau,(i+1)} \leftarrow \mathbf{E}^{\tau,(i)}$ 
12:      end if
13:      Update  $\log(\sigma_{j'}^{(i+1)}) \leftarrow \log(\sigma_{j'}^{(i)}) + (1+i)^{-0.5}(\alpha - 0.44)$ 
14:       $i_{k+1} \leftarrow i_k + 1$ 
15:    end for
16:  end for

```

4.3.5 Summary

Through **Algoithm 5** we define an efficient sampling procedure for the state space $k = \{\theta, \mathbf{E}, \mathbf{E}^\tau, \mathbf{I}\}$ within one single framework. This allows us to infer the exposure state, infection state, exposure and infection timings across the epidemic, the correlates of protection and the antibody kinetics function. We show the ability of this procedure to recover our simulated data displayed in the next section. For this model, we choose a non-informative for the timing of infection given exposure prior: $P_t(E_j^{\tau}) = 1/120$.

4.4 Simulation recovery

After running **Algorithm 5**, we plot the posterior samples, $\hat{\theta}, \hat{\mathbf{I}}, \hat{\mathbf{E}}$, and $\hat{\mathbf{E}}^\tau$ and compare with the simulated parameters.

4.4.1 Exposure state recovery

Algorithm 5 can predict and recover the population-level exposure rates, but struggles on the individual level. For those exposed and infected, the exposure rate is consistently recovered, except when the pre-infection titre is greater than 3.3, in which the boosting is completely attenuated. At these high titres, infection, exposed and not infected, and non-exposure all have the same antibody kinetics (i.e. unchanged

titre throughout the study), and thus, the model cannot differentiate between these exposure-infection states on an individual level. These values are roughly $E[\hat{E}] = 0.5$, implying no preference between a positive or negative state. The model also cannot identify which individuals are exposed and not infected and which individuals are not exposed for all titre values. This is unsurprising, as both these individuals have the same antibody kinetics regardless of pre-infection titre (i.e. unchanged throughout the study). In the case of a correlation of protection with a logistic function, there is some inference on the posterior probability of exposure \hat{E} given pre-infection titre. A low titre values, it is unlikely an individual is exposed and not infected, as the probability of infection given exposure is high. Therefore, the model infers these individuals are unlikely to have been exposed $E[\hat{E}] < 0.5$. This COP influence causes the increasing value of the $E[\hat{E}]$ for each individual as the titre increases in **Figures 7d-e**. Finally, the number of exposed individual is approximately recovered for all six models.

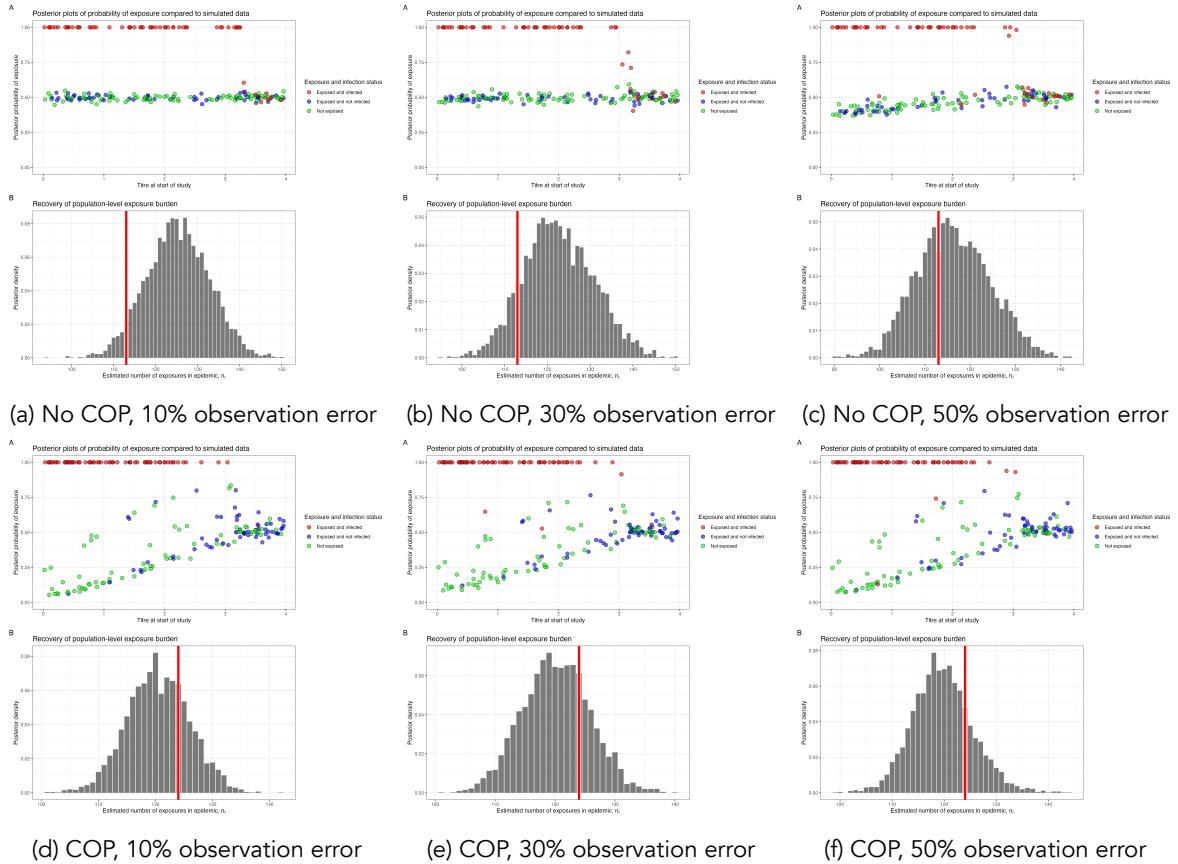


Figure 7: Simulation recovery of exposure status \hat{E} and epidemic curve for two COP models (top: No COP, bottom: logistic COP) and three different levels antibody kinetics variability (10%, 30%, 50%)

4.4.2 Exposure times recovery

Algorithm 5 can recover the exposure times under specific conditions. The posterior of the exposure times for individual j is given by \hat{E}_j^τ and are plotted with the simulated exposure time in **Figures ??** by their exposure status. For those infected, the model can reasonably recover the exposure times for each individual, though as the antibody kinetics variability increases this ability gets poorer (Figures). For those who are exposed and not infected, the model cannot infer the exposure time as there is inference method for this in the likelihood (i.e. exposure time is determined by antibody kinetics, and antibody kinetics remain unchanged for those exposed and not infected). Therefore the inferred exposure time for these individual is the same as the prior entered into the model.

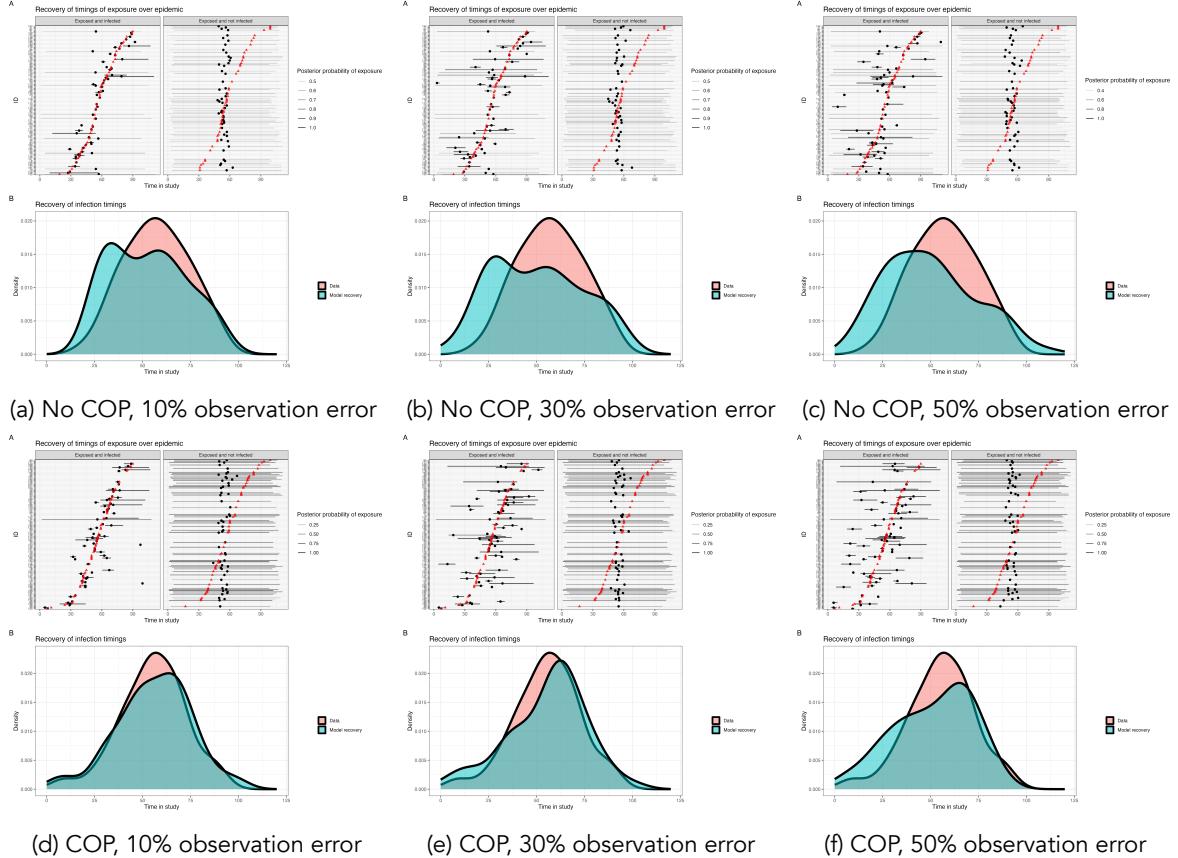


Figure 8: Simulation recovery of exposure and infection timings \hat{E}^τ and epidemic curve for two COP models (top: No COP, bottom: logistic COP) and three different levels antibody kinetics variability (10%, 30%, 50%)

4.4.3 Infection state recovery

We also recover the infection status of each individual from the simulated data. If the set posterior samples of the infection status for individual j is given by \hat{I}_j , then we plot the expectation $\mathbb{E}(\hat{I}_j)$ so we can assess the ability of the algorithm to recover the individual-level simulated infection status (**Figure ??**). As before, we find when the pre-infection titre < 3.3 log titre value that all six models considered can recover the infection status of almost all individuals. When the pre-infection titre is greater than 3.3, the attenuation of boosting for infected individuals causes no meaningful change in the antibody kinetics ($f_{ab}^2(Z, \alpha) = 0$ when $Z > 3.3$). Thus, these individuals' infections are difficult to infer serologically as their titre dynamics are equivalent to independent of their infection status. In our COP model B, we find that including the correlation of protection influences the infection status. As the inferred correlate has a low probability of infection at higher titres, this causes the \hat{I}_j to be more likely to be 0 at higher titre values.

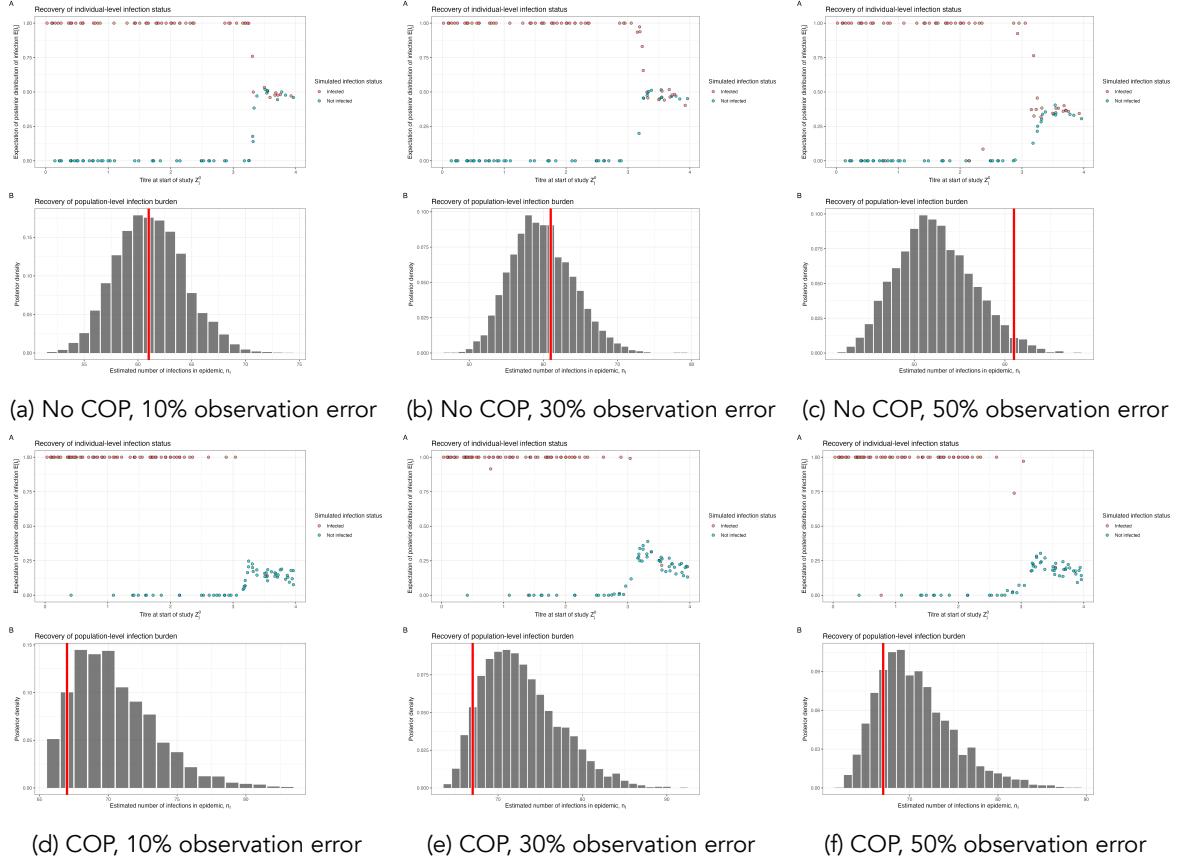


Figure 9: Simulation recovery of the individual infection status, \hat{I}_j , for two COP models (top: No COP, bottom: logistic COP) and three different levels antibody kinetics variability (10%, 30%, 50%)

4.4.4 Correlate of protection

Algorithm 5 performs well at recovering the correlate of protection function $f_{cop}(x, \hat{\theta}_{cop})$, where x is the titre value at infection and where $\hat{\theta}_{cop} = \{\hat{\beta}_0, \hat{\beta}_1\}$ are the posterior samples for β_0 and β_1 . For Model A, we find that the COP curve is recovered, with the simulated line within a 50% confidence interval of the posterior sample (**Figure ??**). For Model B, we find the logistic shape of the COP is recovered in the posterior samples.

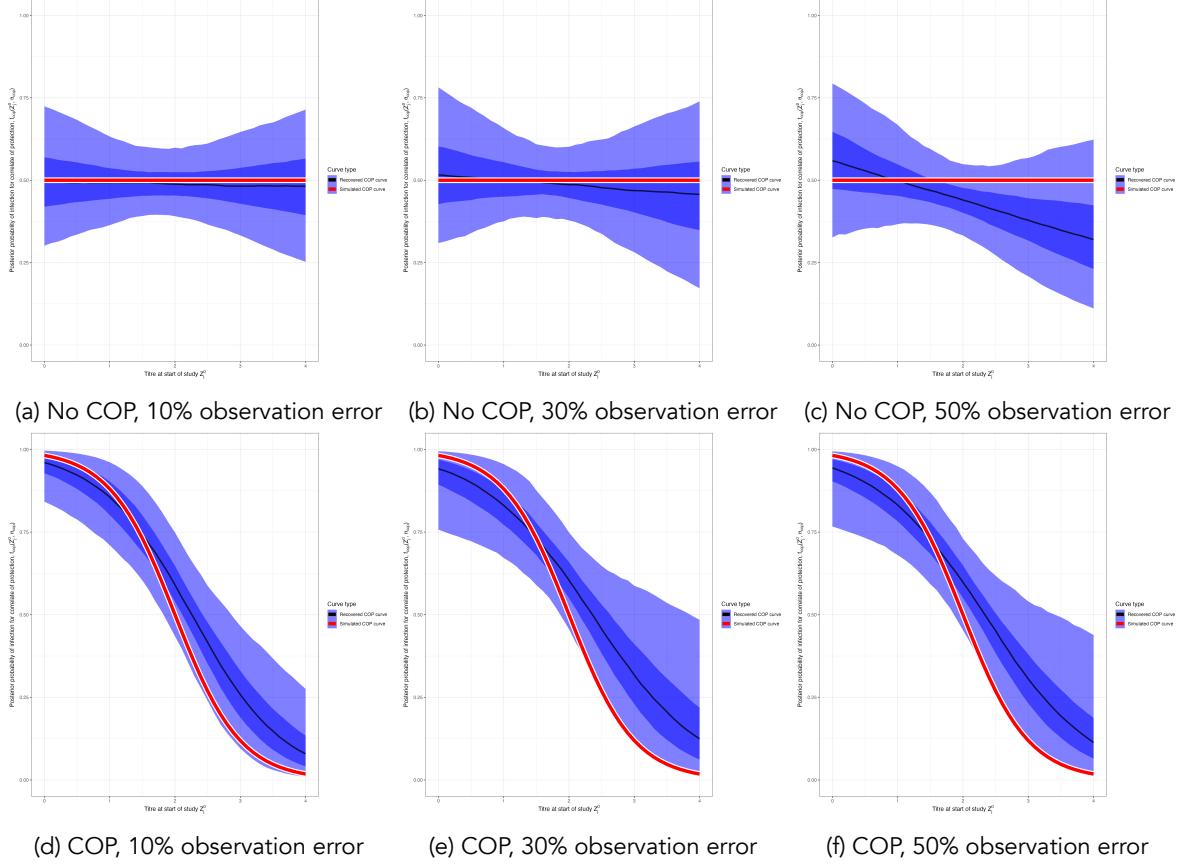


Figure 10: Simulation recovery of the COP function, posterior samples plot $f_{cop}(x, \hat{\theta}_{cop})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).

4.4.5 Antibody kinetics

Algorithm 5 also successfully recovers the simulated antibody kinetics. Let us plot $f_{ab}^1(s, \hat{a}, \hat{b}, \hat{c})$, the posterior predictive distribution for the antibody kinetic boosting, given posterior distributions for \hat{a} , \hat{b} , and \hat{c} . At all three levels of kinetic uncertainty, the antibody kinetics are recovered, though increasing uncertainty weakens the accuracy of the recovered curves compared to the simulated. (**Figure ??**).

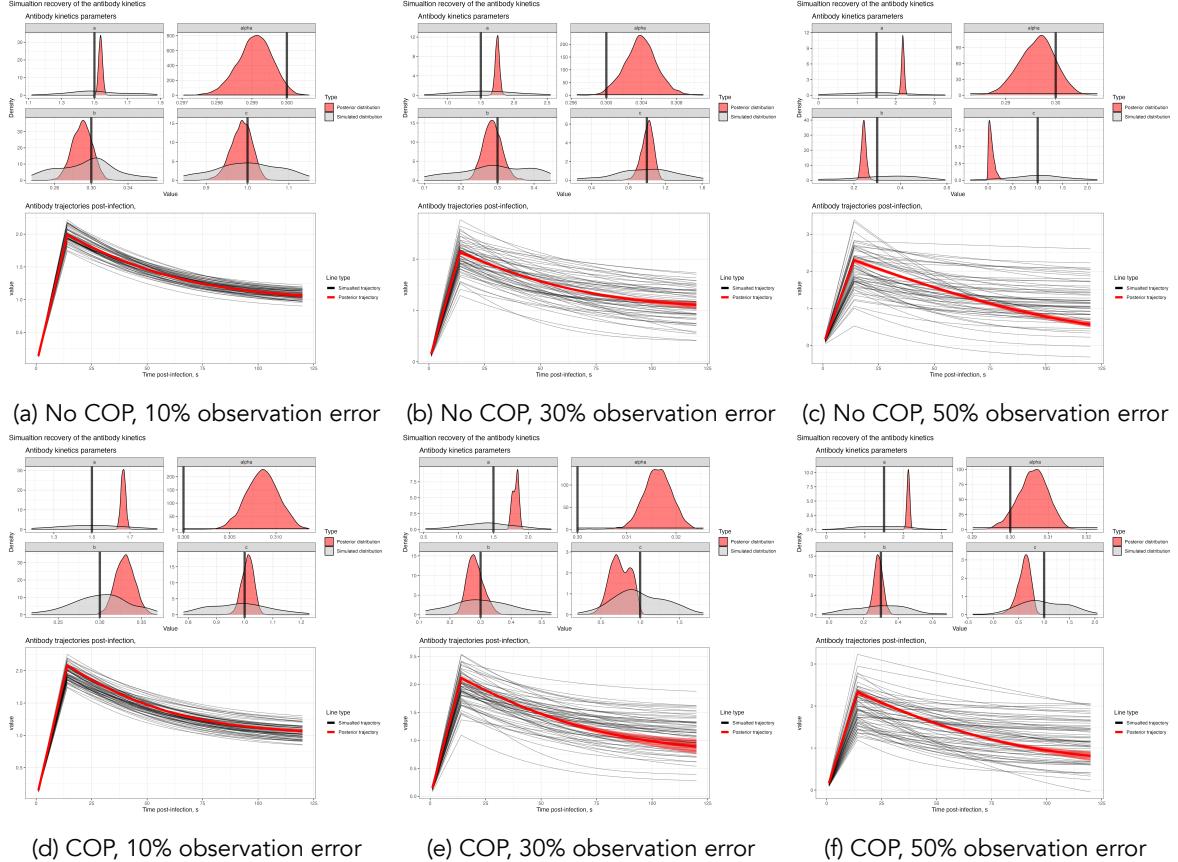


Figure 11: Simulation recovery of the antibody kinetics function with posterior samples plot $f_{ab}^1(s, \hat{a}, \hat{b}, \hat{c})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).

5 Looking forward

In summary, we have shown the ability of **Algorithm 5** to recover the state variables $\{\theta\}$ from simulated serological data. Additionally, we have shown how well the correlation of protection and antibody kinetics functions are recovered, showing that they are well-recovered for all six of our models chosen. The model cannot infer individual-level exposure for those not infected. However, this is unsurprising as the kinetics between these groups are equivalent. The proportion of the population exposed is well recovered, however. At high levels of variability in the antibody kinetics, the ability of the model to infer the exposure time on the individual level weakens and the epidemic curves starts to differ from the simulated curve. Finally, the COP infection and infection status are well-recovered for all models considered, suggesting that their inference is more reliable in the face of high-level individual-level variability in antibody kinetics.

RJMCMC algorithms have been used in infectious disease modelling previously. Hendrick?

These models are incredibly useful for several reasons: Immunobridging?

Extensions in the future

- Add the possibility of inferring multiple exposures for an individual
- Add hierarchical effects to antibodies kinetics and correlates of protection
- Add inferring for multiple biomarkers and antigenically varied pathogens to improve inference
- Methods development: parallel tempering? Accessibility of method to others

This document has provided details of the theoretical underpinning and implementation of a reversible jump mcmc algorithm, which can infer important epidemiology and immunological information from individual-level serological data. On the individual level, it can infer the exposure status, infection status, infection timelines, the antibody kinetics and the correlate of protection for each individual.

To conclude, this documents provides a walkthrough of how to implement a reversible jump algorithm to infer serological data. We hope this technique will be useful for inferring epidemiological information in a pathogen-agnostic setting, particularly pathogens where intense surveillance is challenging. We also hope this document sheds light on a mathematically complex but powerful inferring tool and encourages others to implement similar algorithms in other areas of health science which require the exploration of multidimensional model spaces.

6 Appendix

6.1 Adaptive Proposal Distribution

I use an adaptive proposal distribution $q_\theta(\theta)$ to sample the parameter space θ . The adaptive metropolis hasting algorithm provides systematic method for modifying the shape of the proposal distribution based on the accepted steps of the current markov chain, allowing for more efficient mixing of chains. That is the $q_\theta(\theta_i) = N(\theta_i, \Sigma_i(\theta_i))$ follows a Gaussian distribution. To provide a reasonable estimate for the covariance matrix Σ_i , the Markov chain runs for an initial number of steps (T_{init}) from a truncated multivariate normal proposal distribution with a covariance matrix, I_s , whose entries are calculated using the upper and lower bounds of the support of the priors $[s_0^k, s_1^k] \in \mathcal{S}$, through $i_{k,k} = (s_1^k - s_0^k)/\zeta$ and $i_{i,j} = 0$ otherwise, where ζ is a scaling factor.

Problematically, the proposal distribution using the updated covariance matrix, Σ_i , is no longer memoryless, and therefore chain may no longer converge to the correct stationary distribution. To overcome this problem, the proposal distribution must also sample from a non-adaptive multivariate Gaussian distribution modified to ensure that changes to the covariance matrix diminish over time. Further, to improve

chain mixing and to optimise convergence rates, I include adaptive scaling factors, λ_i and M_i for the initial non-adaptive and adaptive proposals, respectively, whose magnitude diminishes with the number of steps in the chain. The adaptive scaling factor for the non-adaptive proposal distributions stops once the model starts sampling from the adaptive proposal distributions. Overall, the combined non-adaptive and adaptive proposal distributions for the adaptive Metropolis Hastings is given by

$$\begin{array}{c|c|c} i & i \leq T_{init} & i > T_{init} \\ \hline q(\cdot | \theta_i) & \mathcal{N}(\theta_i, \exp(\lambda_i)I_s; \mathcal{S}) & \begin{array}{ll} \mathcal{N}(\theta_i, \Sigma_i; \mathcal{S}) & \text{with probability } \beta, \\ \mathcal{N}(\theta_i, \exp(\lambda_{t_{init}})I_s; \mathcal{S}) & \text{with probability } 1 - \beta \end{array} \end{array} \quad (31)$$

where $\Sigma_t = \exp(M_i)\Gamma_i$ and M_i , λ_i and Γ_i are updated iteratively through the stochastic approximation algorithm:

$$\begin{aligned} \lambda_{i+1} &= \lambda_i + \gamma_1(i)(a(\theta_i, \theta^*) - 0.234) \\ M_{i+1} &= M_i + \gamma_2(i)(a(\theta_t, \theta^*) - 0.234) \\ \mu_{i+1} &= \mu_i + \gamma_3(i)(\mu_t - \theta_t) \\ \Gamma_{i+1} &= \Gamma_i + \gamma_4(i)[(\theta_i - \mu_{i+1})(\theta_t - \mu_{i+1})^T - \Gamma_t] \end{aligned}$$

where $\gamma_i(t)$ are gain factors. Note when $i > T_{init}$ up stop updaing λ_i .

In our implementation, we define $\theta_{i,adapt} = \{M_i, \mu_i, \Gamma_i, \lambda_i\}$, and choose values, $\beta = 0.05$, $\zeta = 100$, $\lambda_0 = \log(0.1^2/|\theta_i|)$, $M_0 = \log(2.382^2/|\theta_i|)$, $\mu_0 = \pi_0$, $\Gamma_0 = I_s$, and $\gamma_x(i) = (1+i)^{-0.5}$ for all x .