Modelling serological data using a reversible jump mcmc algorithm

David Hodgson

January 10, 2024

1 Overview of serological modelling

1.1 Serological sampling

Serological samples, such as blood or serum, can be used to detect the presence of biomarkers produced by the immune system in response to a specific pathogen. This allows researchers and healthcare professionals to deduce crucial information about the epidemiological of the pathogen on the individual and population-level which may otherwise be missed by active surveillance systems.

On the individual-level, serological samples can be used to determine if an individual has been previously infected with a particular infectious agent, even if the symptoms have resolved or were asymptomatic. After measuring the quantity of antibodies to a specific pathogen or pathogen, infection is usually inferred using an antibody threshold or by a threshold fold-rise between a pair of samples. In addition, researchers can study the specific immune responses generated by individuals during infection and see how they change according to certain host factors. This information can lead to a better understanding of the immune system's ability to combat various pathogens, aiding in the development of new treatments and vaccines. Studying Disease Dynamics: Serological data can provide insights into the transmission dynamics of infectious diseases. By analyzing changes in antibody prevalence over time, researchers can track the spread of a disease, identify hotspots, and better understand how pathogens move through populations.

On the population-level, serological samples which are representative of a population, researchers can by used to estimate the prevalence of a specific infectious disease within that community. This information is invaluable for public health planning, resource allocation, and understanding the burden of a disease in a given region. In addition, serological tests can help assess the level of immunity within a population. This information is essential for designing vaccination campaigns, identifying vulnerable groups, and evaluating the effectiveness of vaccination programs. It also aids in making informed decisions during disease outbreaks. Finally, serological samples can be used to evaluate the effectiveness of vaccines by measuring the presence and levels of specific antibodies in vaccinated individuals. This data helps determine if a vaccine is providing adequate protection against the targeted pathogen and if booster shots are necessary.

Serological samples play a pivotal role in understanding infectious diseases by providing information about past infections, disease prevalence, immunity levels, vaccine efficacy, and disease dynamics. They play an increasingly important role in public health efforts to combat and control infectious diseases. However, inferring infection requires a derivation of a abosolute or relative thresholdhold value however this is often determined heuristics (e.g. flu), but is highly variable between individuals. Consequently, better understanding of the kinetics of antibody trajectories post vaccination and infection can provide a better understanding of the heuristics used to infer infections from serological samples

1.2 Antibody kineitcs

Modeling antibody kinetics involves using mathematical and statistical techniques to understand dynamics 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. Key components of such models include: This process is essential for understanding how antibodies develop, peak, decline, and potentially provide protection against infectious diseases. Several attemps have been made to understand the kinetics of antibody kinetics, typically it follows a three stage process

- Initial Response: Modeling often starts 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 defense against the antigen.
- vPeak Antibody Level: The models track the 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: Antibody kinetics models also account for the decline in antibody levels following the peak response. Antibodies have a finite lifespan in the bloodstream, and their concentration gradually decreases as the pathogen is cleared or the vaccine antigen wanes.

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. Models help in understanding how changes in antibody levels impact disease transmission dynamics. This is vital for predicting and controlling outbreaks.

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 levels of immunity. Models allow researchers to evaluate the effectiveness of vaccines by comparing predicted antibody responses to observed data. This assessment aids in vaccine development and refinement.

In conclusion, modeling antibody kinetics is a valuable tool for understanding the dynamics of immune responses to infections and vaccinations. It provides insights into immunity duration, vaccine efficacy, and strategies for disease prevention and control, ultimately contributing to better public health outcomes.

1.3 Correlates of protection

Determining correlates of protection for infectious diseases is crucial for vaccine development and public health planning, but it can be challenging due to several factors. Here, we'll explore both the importance and the difficulties associated with identifying these correlates. 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 that can effectively induce the required immune response.

Correlates also serve as critical benchmarks in assessing the effectiveness of vaccines during clinical trials. They allow researchers to measure whether a vaccine candidate can generate the necessary immune response and provide protection against the disease. By identifying correlates of protection informs vaccination strategies, including dosing schedules and target populations. It ensures that vaccines are administered optimally to achieve and maintain immunity, thus preventing outbreaks. Finally, understanding correlates of protection aids in devising strategies for controlling and eventually eradicating infectious diseases. By targeting specific immune responses, interventions can be more effective.

Determining whether someone has been exposed to an infection but protected due to existing antibody levels is key in determining correlates of protection. Those who have been infected have also been exposed, but those who do not acquire infection detectable by antibodies, or the infection is aborted by other immune processes, are particularly difficult to detected with out without intense contact tracing which is challanging and difficult to roll-out on a large scale, or intense immune profiling. Further diseases may require unique approaches, and not all diseases have well-established correlates. Existing correlates have been determined through Challenge Studies: In controlled challenge studies, volunteers are deliberately exposed to the pathogen after vaccination. Researchers monitor their immune responses and assess whether they remain protected or develop the disease. This approach allows for a direct evaluation of correlates of protection. However these are expensive and determine.

In conclusion, determining exposure to infections in a population is vital for public health efforts, but it comes with significant challenges, including asymptomatic cases, testing limitations, and the complexities of immune responses. Overcoming these challenges requires a combination of improved testing strategies, data collection, and the continued development of diagnostic tools to ensure accurate and timely identification of exposed individuals.

1.4 Determining correlates of protection and antibody kinetics in a single framework

To overcome the issues of biasing antibody kinetics due to lack of informatin on exposure, we present a single framework which takes individual-level serological sample data and uses changes in antibody titres over time to determine i) the subsequent antibody kinetics of these infection individuals and ii) the correlate of protection preventing exposued individuals from becoming infected. The intermeidate process which allows this is a process which determines exposed and individuals who are infected with the virus, into a single mathematical framework we can properly quantifythe uncertainty in the framework by allows interdependencies between these modules (i.e. antibody kinetics, those infected) and provide more accurate understanding of the serological inference with out the need for pre-determined heureistics.

1.5 Simulated data

To explain the framework we simulate serological data using the `serosim`[ref] R package. We simulate continuous epidemic serosurveillence (CES) cohort data, but we also provide information in the appendix for pre- and post-pandemic serosurvey (PPES) cohort data. CES data 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 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. To model an even epidemic peak, we simulate the exposure rate for each individual from a normal distribution, $N(60, 20)$.

The correlate of protection is the probability of infection given a titre value at exposure. Two different sets of data are simulated with two different correlates of protection; one is uniform for all titres at exposure and thus represents no correlation of protection. The second follows a logistic distribution of the form:

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

where $\beta_0 = 2$ and $\beta_1 = 2$ in the simulated data. This represents a pathogen for which higher antibody titres are associated with higher levels of protection from 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=piont value as defined in X[ref]. The formula for this biphasic trajectory is given by Equation 2

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

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 are dependent on pre-existing titre value, with higher pre-existing values seeing attenuated dynamics relative to lower pre-existing tire values. The titre dependent boosting is assumed to follow a linear decay truncated at 0, that is, given a titre value z, $f_{tp}(z,\alpha) = \max(1-\alpha z, 0)$. where $\alpha = 0.3$ in the simulated data. Therefore the value to the value depends on the time t and titre value at infection z;

$$f_{ab}(s,z,a,b,c,\alpha) = F_{ab}(s,a,b,c) f_{tp}(z,\alpha) \quad (3)$$

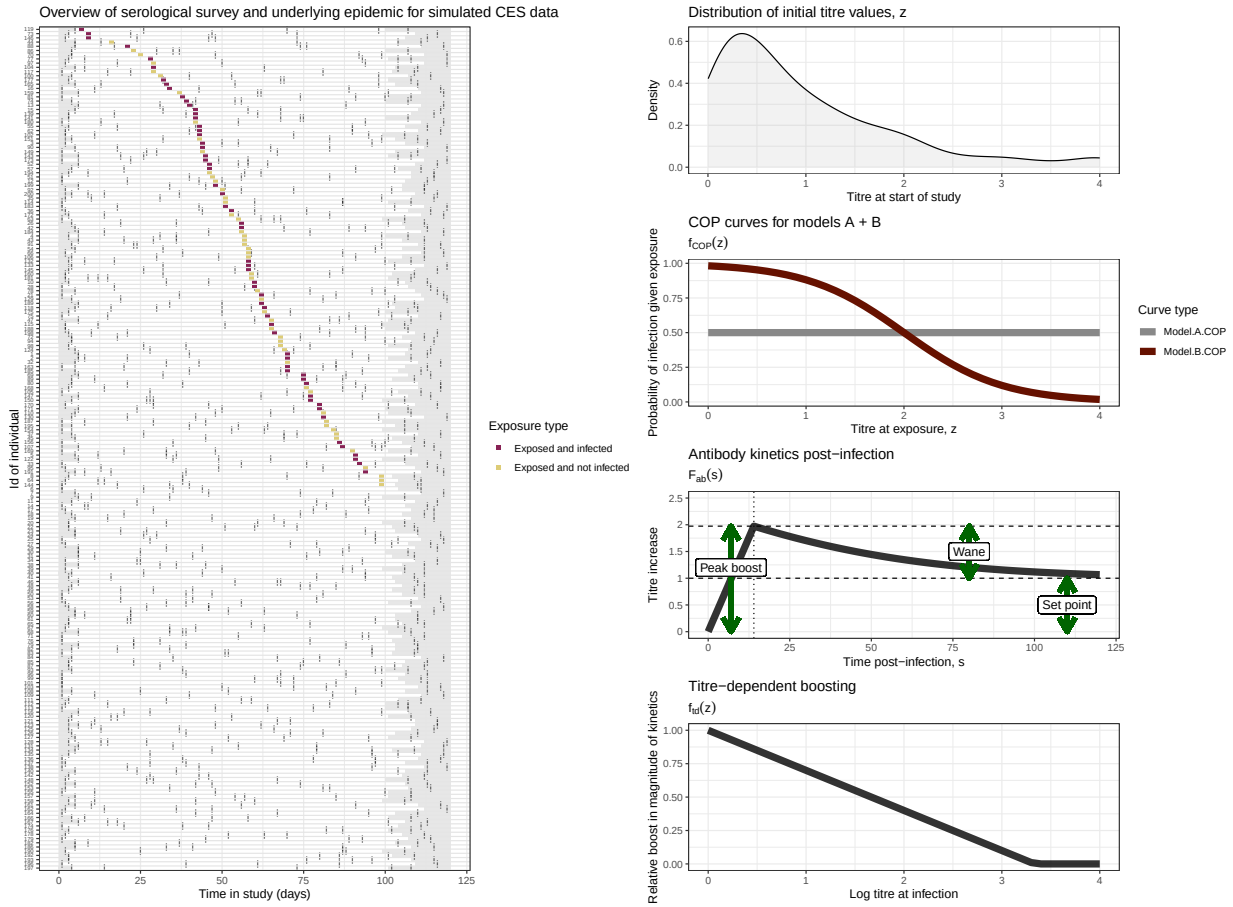

Figure 1: Schematics showing the simulated data structure from `serosim`[ref]

To model individual-level in antibody kinetics, capturing both indiviudal hetergeneity in repsonses nad. measurement uncertaity, we simulate responses with for a, b, c, α from a normal distributions centered at where X is chosen. We simulate three llevels of uncetainty, 0.01, 0.2, 0.5 for each correlate of protection model, giving six datasets in total. A schematic showing how the variance influences the trajectories of a variability of the antibody kinetics is shown in Figure X.

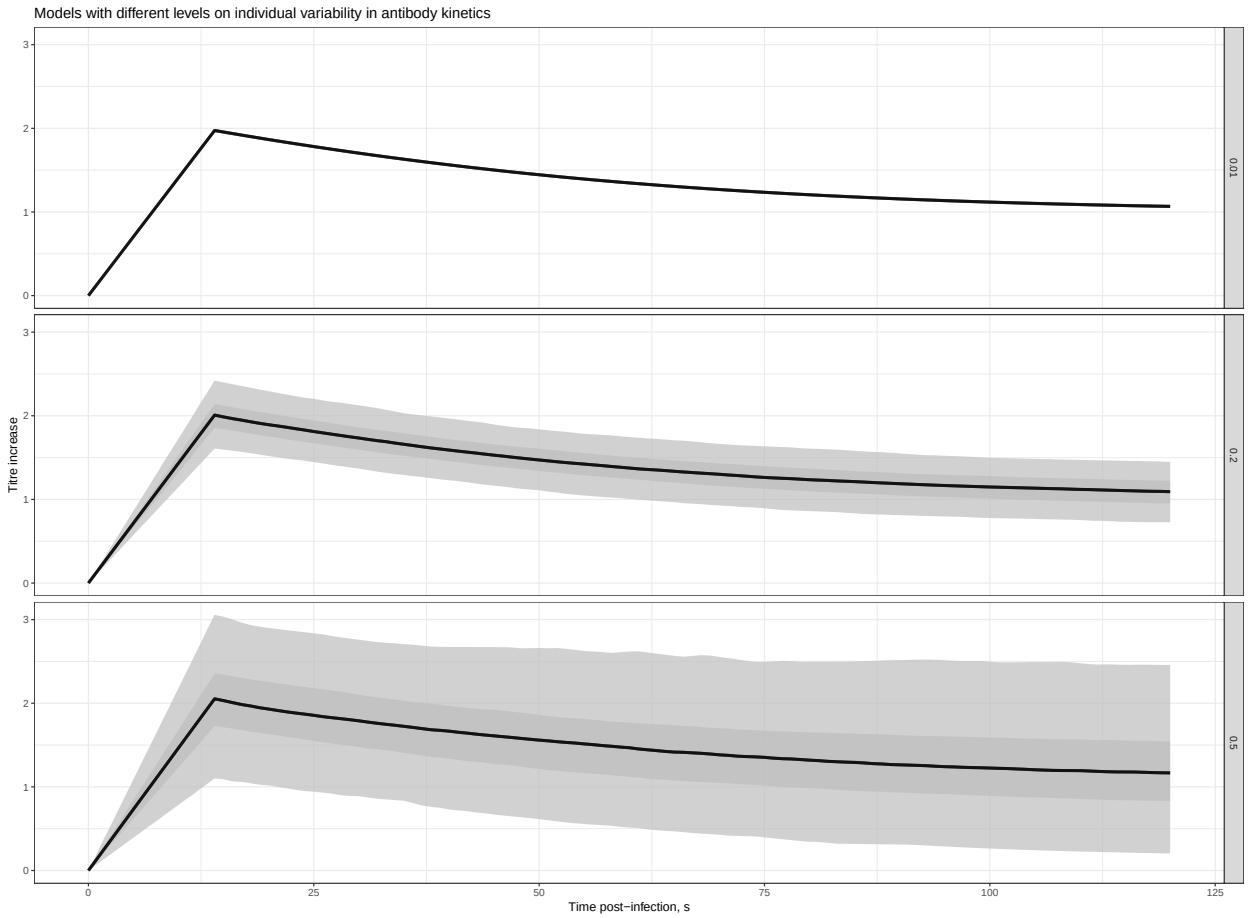

Figure 2: Schematics showing three levels of individual-level uncertainty and the impact on the variability of antibody kinetics.

For these six simulated data sets, information on the number exposed and infected is given in Table X. We will fit the to-be-described methods to these datasets and explore how the correlation of protection and level of variability in in antibody kinetics impacts the framework's ability to recovery simulated data.

2 Inference with known exposure status

We now provide a methodological walkthrough for how the reversible-jump mcmc serological (RJMCMC) framework works. First, we will show simulation recovery in a pedagogical case in which we assume the time of exposure and exposure status is known. Though this is not possible in practice, this example will help explain how the MCMC sampler works without having to describe the more complex RJMCMC. We show that the correlate of protection and antibody kinetic s are well recovered. In the second section, we show assume that exposure status and time is not known for each indivdual, and describe how the RJMCMC framework works. We find that the timing and exposure, infection status, correlation of protection and antibody kinetics are recovered for our simulated data.

2.1 Mathematical representation

For each individual, i, has three states that must be inferred, E_i, whether they are exposed over the study period, E_i^t the timing of their exposure, I_i whether they are infected (Table **??**).

Table 1: Inferrable epidemiological states

Symbol	Description	Values
E_i	Exposure status of individual i	Binary value. 0: Not exposure, 1: exposed.
E_i^t	Time of exposure for individual i	Continuous value between start of study ($T = 0$) and values of study ($T = T_{end}$)
I_i	Infection status of individual i.	Binary value. 0: Not infected, 1: Infected.

For example, a person who is not exposed has values E = 0, E = NA, and I is always 1. For E = 1, E = 60 days (after the start of the study), and I is either 0 or 1, depending on the protection correlation. In this example, we know the value of E_i and E_i^t for each individual. Thus we need only infer the status of I.

We first define several functions which will help us calculate our likelihood. First, we assume that the model predicted antibody titre at time t in the study ($X_{i.t}$) can be derived given the infection status I_i and timing of exposure E_i^t. If a person is not infected, then their starting titre value (Z_i^0) remains unchanged from the state 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 Eq[]. The deterministic function for calculating this value is given by

$$X_{i,t} = P_{ab}(I_i, E_i^t, a, b, c, \alpha, Z_i^0) = \begin{cases} Z_i^0, & \text{If } I_i = 0 \text{ or } E_i = 0 \text{ or if } E_i^t > t \\ Z_i^0 + f_{ab}(t - E_i^t, Z_i^0, a, b, c, \alpha), & \text{Otherwise} \end{cases} \tag{4}$$

Second, we define a likelihood function for the correlation of protection. For an individual i, with $E_i = 1$, the correlate of protection given exposure at time t with titre value $X_{i.t}$, given by a Bernoulli distribution with the probability is given by X

$$P_{cop}(I_i \mid E_i = 1, X_{i.t}, \beta_0, \beta_1) = f_{cop}(X_{i.t}, \beta_0, \beta_1)^{I_i}(1 - f_{cop}(X_{i.t}, \beta_0, \beta_1))^{1-I_i} \tag{5}$$

Finally, we define an observational model to capture variability between hosts and measurement error. Assuming that the model predicted titre value at time t for individual i is given by, $X_{i.t}$ and the serologicla antibody body at the same time point is given by $Z_{i.t}$, we assume the measrement error follow a normal distribution and dervied the following likliehood.

$$P_{obs}(Z_{i,t} \mid X_{i.t}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z_{i,t} - X_{i.t})^2}{2\sigma^2}} \tag{6}$$

where σ is fitted in the model. We let $\theta = \{a, b, c, \alpha, \beta_0, \beta_1\}$ for clarity henceforth.

2.2 Posterior distribution via Bayes rule

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

2.2.1 Case 1. $E_i = 0$

In this case the value of the timing of exposure is not applicable and the individual cannot have been infected (i.e. $E_i^t = \text{NA}, I_i = 0$). The likelihood for this individual with serological samples taken at times $t \in T$ is therefore equivalent to:

$$L_{E_i=0}(Z_i|\theta) = \prod_{t \in T} P_{obs}(Z_{i,t}|Z_i^0, \sigma) \tag{7}$$

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

2.2.2 Case 2. $E_i = 1$

In this case, the value of the timing of exposure is taken from a user-defined probability distribution approximated to the epidemic curve (f_I), and the infectious status is determined by the correlate of protection function (f_{cop}). (i.e. $E_i^t = \text{NA}, I_i = 0$). The likelihood for this individual with serological samples taken at times $t \in T$ and infection time τ is therefore equivalent to:

$$L_{E_i=1}(Z_i|I_i, \theta) = \prod_{t \in T} P_{obs}(Z_{i,t}|X_{i,t}, \sigma) P_{cop}(I_i \mid X_{i.t}, \theta) \tag{8}$$

where $X_{i,t} = P_{ab}(I_i, E_i^t, a, b, c, \alpha, Z_i^0)$.

2.2.3 Total likelihood

If $N_{E=0}$ and $N_{E=1}$ is the set of individuals which are not exposure and exposed respectively, then the total likelihood can be written

$$L(Z|I, N_{E=0}, N_{E=1}, \theta) = \prod_{i \in N_{E=0}} L_{E_i=0}(Z_i|\theta) \prod_{i \in N_{E=1}} L_{E_i=1}(Z_i|I_i, \theta) \tag{9}$$

2.3 Metropolis-Hasting algorithm

2.3.1 Overview

The Metropolis-Hastings 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 distribttion $p(x)$. The idea of the MH is to define a Markov chain over possible values in such a way that the stationary distribution of the Markov chain is in fact $p(x)$. That is, the resulting Markov chain from MH generates a sequence of values, denoted $\{\theta_1, \theta_2, \ldots, \theta_n\}$, such that as $n- > \infty$ we can guarantee that $theta_i \sim p(x)$. The MH algorithm is a two part process, first a proposal value θ^* is sampled from a proposal distribution, Q, centered at the current chain step θ, $\theta \sim Q(\theta)$. This proposal is then accept according to a acceptance ratio whichis the ratio of the posterior distribution evaluated at the newly sampled chain and the posterior distribution sampled at the current chain. A general algorithm is given in Algorithm **??**..

Algorithm 1 Generic Metropolis-Hastings Algorithm

Initialize the chain with an initial state θ_0
for $i = 1$ to N **do**
    Generate a candidate state θ' from the proposal distribution: $\theta' \sim Q(\theta' \mid \theta_i)$
    Compute the acceptance ratio:

$$\alpha = \min\left(1, \frac{P(\theta')}{P(\theta_i)} \cdot \frac{Q(\theta_i \mid \theta')}{Q(\theta' \mid \theta_i)}\right)$$

    Generate a uniform random number u from the interval $[0, 1]$
    if $u \leq \alpha$ **then**
        Accept the candidate state: $\theta_{i+1} \leftarrow \theta'$
    else
        Reject the candidate state: $\theta_{i+1} \leftarrow \theta_i$
    end if
end for

2.3.2 MH for serological inference with known exposure

In our datasets we wish to infer θ, and infectious statuses I_i within one framework algorithm. For the proposal distribution we define separate proposal distribution for each I_i and θ, such that $Q(\theta) = q()q()$. The proposal distribution for $q(\theta)$ comes from an adaptive metropolis hasting algorithm with XX. The proposal distribution for I, which is a discrete space, is performed by choosing a individual i such that $E_i = 1$ at each time step, then sampling a value according to a betabinoamial distribution. These prosposal are both symmetric and thus cancel out in the acceptance ratio. Consequently, we edit the algorithm to that as given in Algorithm **??**.

Algorithm 2 Generic Metropolis-Hastings Algorithm

Initialize the chain with an initial state θ_0
for $i = 1$ to N **do**
    Generate a candidate state θ' from the proposal distribution: $\theta' \sim Q(\theta' \mid \theta_i)$
    Generate a candidate inidividual i and then candidate state I_i' from the proposal distribution: $\theta' \sim Q(\theta' \mid \theta_i)$
    Compute the acceptance ratio:

$$\alpha = \min\left(1, \frac{P(\theta')}{P(\theta_i)}\right)$$

    Generate a uniform random number u from the interval $[0, 1]$
    if $u \leq \alpha$ **then**
        Accept the candidate state: $\theta_{i+1} \leftarrow \theta'$
    else
        Reject the candidate state: $\theta_{i+1} \leftarrow \theta_i$
    end if
end for

2.4 Implementation

We code the algorithm manually in cpp with an R interface. We run the algorhtm for XX steps, with XX burn-in and for 4 chains. The starting value is taken from the prior distributions of the for θ and I respectively. We initialise the adaptive covariance by running with an identity matix with each parameter scales according for 1,000 steps then sample from the adaptive scheme as in XX.

2.5 Simulation recovery

After running the algorithm we show the ability of the metropolis hasting algorithm to recover the simulated data highlight in section X.

2.5.1 Antibody kinetics

2.5.2 Correlate of protection

2.5.3 Infection recovery

2.5.4 Summary

For known exposure we have shown this MH algorithm is able to recovery to correlate of protection, infection status, and antibody kinetics for varying correlates of protection models and individual-level variability. In practice this algorithm is unlikely to be useful as the exposure state is not know. In the next section we will expand on this algorithm for the case when exposure status is not known throughout the serosurvey.

3 Inference with unknown exposure status

3.1 Overview

In the case where the exposure status is unknown, we must now infer two more states per individual, their exposure state and the time of exposure given they are exposured. This in the case where $E_i = 0$ the likelihood is as derived before, however in the case where $E_i = 1$ the likelihood contains an additional term to determine the probability of infection and probability of exposure and

$$L_{E_i=1}(Z_i|I_i,\theta) = \prod_{t\in T} P_{obs}(Z_{i,t}|X_{i,t},\sigma)P_{cop}(I_i \mid X_{i.t},\theta) \tag{10}$$

where $X_{i,t} = P_{ab}(I_i, E_i^t, a, b, c, \alpha, Z_i^0)$.

given our combined likelihood is as before, we run into an issue when sampling between different exposure groups, the likelihood above will have more parameters if there are more exposures and thus the detailed balance condition of the mh algorithm now fails. Therefore regardless of out proposal distribution we cannot use the existing algorithm highlighted in X.

3.2 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.

3.2.1 Mathematical overview

This section follows the notations and explaination given in Section 3.3.

Let $\{\mathcal{M}_k : k \in \mathcal{K}\}$ denote a collection models with different dimensions. Let $\mathcal{M}_k$ have it's own parameter space $\Theta_k \in R$. A full Bayesian model for $\mathcal{M}_k$ can be written

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

where $p(k)$ is the prior probability that model $\mathcal{M}_k$ is chosen, $p(\theta_k|k)$ is the prior distribution for parameter θ_k, and

3.3 Application of RJMCMC to serological data

Instead let us consider the space of all possible exposure combinations and extract a value E. Technically the reversible jjump can be used to flip between any of these states but the correction factor part of the rJMCMC becomes complicated. Instead we will consider a birth-death-sampling process which simplifies the RJMCMC framework considerably. This allows small incremental changes between exposure states simplyfying the mathematical resolve.

Birth process, this is where we have an one extra E in comparison to before. to jump between these models the correction term becomes XX.

Death process, this is where we have an one extra E in comparison to before. to jump between these models the correction term becomes XX.

Sampling process, as befrore. Detailed balance condition is met and we sample from the proposal for I and theta as before.