1 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 $\mathbf{E}^\tau$) 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.

1.1 Mathematical representation of framework

Let the binary vector $\mathbf{E} = \{E_1, E_2, \ldots, E_M\}$, 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 $\mathbf{E}^\tau = \{E_1^\tau, E_2^\tau, \ldots, E_{n_{\mathbf{E}}}^\tau\}$ and $\mathbf{I} = \{I_1, I_2^\cdot \ldots, I_{n_{\mathbf{E}}}\}$ 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 $E_j = 0$ 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 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 **??**. The deterministic function for calculating $X_{j,t}$ value is given by

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

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 **??**. The PDF of this likelihood is given by Equation **??**.

$$P_{cop}(I_j \mid Z_j^0, \theta_{cop}) = f_{cop}(Z_j^0, \theta_{cop})^{I_j}(1 - f_{cop}(Z_j^0, \theta_{cop}))^{1-I_j} \tag{2}$$

$\theta_{cop} = \{\beta_0, \beta_1\}$. 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 **??**.

$$P_{obs}(Z_{j,t} \mid X_{j,t}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\left(\frac{(Z_{j,t} - X_{j,t})^2}{2\sigma^2}\right)} \tag{3}$$

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.

1.2 Posterior distribution via Bayes rule

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

1.2.1 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:

$$L_{E_j=0}(Z_j|\theta) = \prod_{t \in T_j} P_{obs}(Z_{j,t}|Z_j^0, \sigma) \tag{4}$$

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

1.2.2 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:

$$L_{E_j=1}(Z_j|I_j,\theta) = \prod_{t\in T_j} P_{obs}(Z_{j,t}|X_{j,t},\sigma)P_{cop}(I_j \mid Z_j^0,\theta_{cop}) \tag{5}$$

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

1.2.3 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

$$L(\mathbf{Z}|\mathbf{I},\theta) = \prod_{j\in\mathbf{E_0}} L_{E_j=0}(Z_j|\theta) \prod_{j\in\mathbf{E_1}} L_{E_j=1}(Z_j|I_j,\theta) \tag{6}$$

1.2.4 Prior distributions

We choose prior distributions for each parameter $\pi(\theta)$. **Table ??** 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 rise from products of Bernoulli trials**??** as $\text{BetaBinomial}(n_{\mathbf{I}}|n_{\mathbf{E}}1, 1) = 1/n_{\mathbf{E}}$ for all $0 \leq n_{\mathbf{I}} \leq n_{\mathbf{E}}$.

1.2.5 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

$$P(\theta, \mathbf{I}|\mathbf{Z}) \propto \mathcal{L}(\mathbf{Z}|\mathbf{I},\theta)\pi(\theta)\pi(\mathbf{I}) \tag{7}$$

1.3 Metropolis-Hasting algorithm

1.3.1 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 distribttion $P(x)$. 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 $\{x_1, x_2, \ldots, x_n\}$, such that as $n \to \infty$ we can guarantee that $x_i \sim P(x)$. To do this, we uniquely define the Markov chain by it's transition probabilities from x to x', $F(x', x)$, that must satisfy the detailed balance condition:

$$F(x' \mid x)P(x) = F(x \mid x')P(x') \tag{8}$$

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 $\alpha(x, x')$:

$$F(x' \mid x) = q(x'|x)\alpha(x, x') \tag{9}$$

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

$$\alpha(x, x') = \min\left(1, \frac{P(x')}{P(x)} \cdot \frac{Q(x \mid x')}{Q(x' \mid x)}\right) \tag{10}$$

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

Algorithm 1 Generic Metropolis-Hastings Algorithm

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

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

    Sample $u \sim \mathcal{U}(0, 1)$
    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

1.3.2 MH for serological inference with known exposure

In our model, we wish to sample from the posterior density function given by **Equation ??**, 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 $Q(\theta, \mathcal{I}) = q_\theta(\theta)q_I(\mathbf{I})$. 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 $\theta \sim q_\theta(\cdot|\theta^{(i)}, \psi^{(i)}_{adapt})$. This proposal is a multivariate normal distribution with an adaptive covariance matrix, which is defined by the set of parameters $\psi^{(i)}_{adapt}$, 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 $I_j' \sim q_I(\cdot|I_j^{(i)}) =$ Bernoulli(0.5). Therefore the proposal for $q_I(\mathbf{I}'|\mathbf{I}) = 1/n_{\mathbf{E}}0.5$ for all j. Both of these proposals $q_\theta\left(\theta|\theta^{(i)}, \psi_{adapt}^{(i)}\right)$, $q_I(\mathbf{I}'|\mathbf{I})$ are both symmetric and thus cancel out the acceptance ratio (**Equation ??**). Further, the prior distribution $\pi(\mathbf{I}) = 1/n_{\mathbf{E}}$ for all $0 \leq n_{\mathbf{I}} \leq n_{\mathbf{E}}$, 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 ??**).

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

Initialize the chain with an initial state $\theta^{(0)}$ from the priors $\pi(\cdot)$ and $I_j^{(0)} \sim$ Bernoulli(0.5) for all $1 \leq j \leq M$ individuals to intialise $\mathbf{I}^{(0)}$, and initialise $\psi_{adapt}^{(0)}$.
for $i = 1$ to N **do**
    Generate a candidate state $\theta' \sim q_\theta\left(\theta^{(i)}, \psi_{adapt}^{(0)}\right)$
    Generate a candidate individual $j \in \mathbf{E_1}$, then a candidate state $I_j' \sim$ Bernoulli(0.5) to propose $\mathbf{I}'$
    Compute the acceptance ratio:

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

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

1.4 Implementation

Algorithm ?? 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.

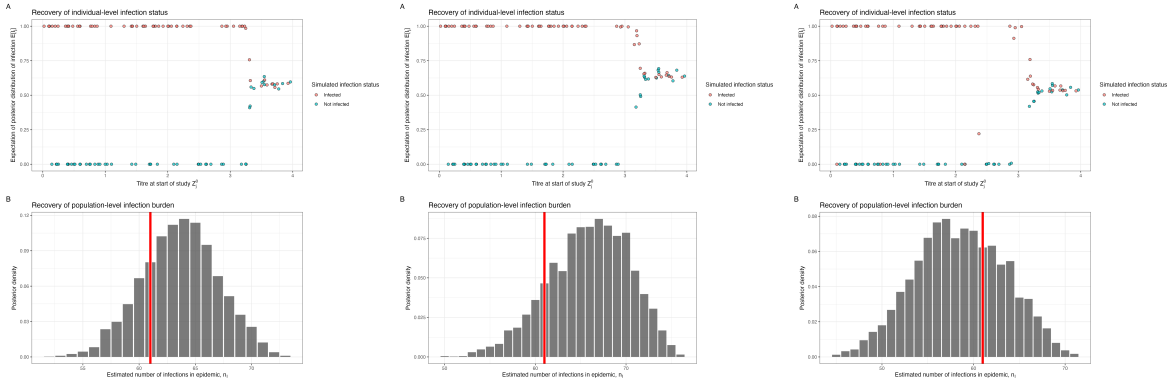
1.5 Simulation recovery

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

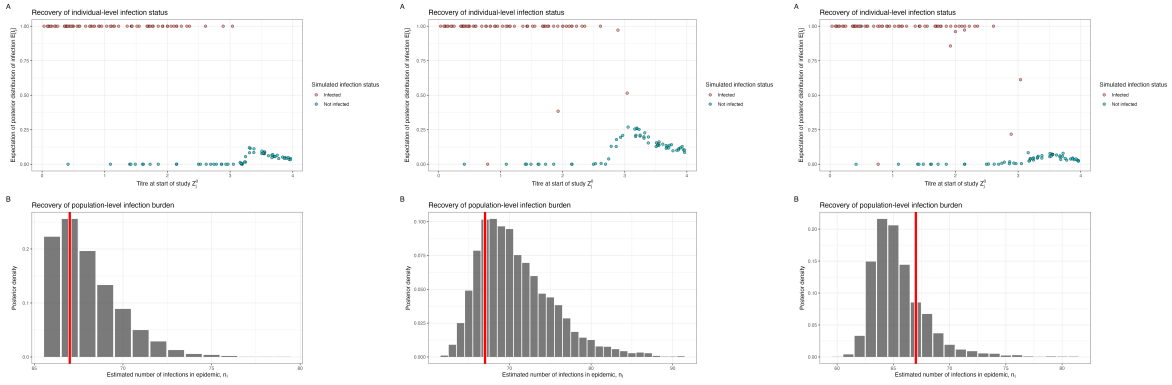
1.5.1 Infection recovery

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 $\mathbb{E}(hatI_j)$ so we can assess the ability of the algorithm to recover the individual-level simulated infection status (**Figure ??**). 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.



(a) No COP, 10% observation error (b) No COP, 30% observation error (c) No COP, 50% observation error

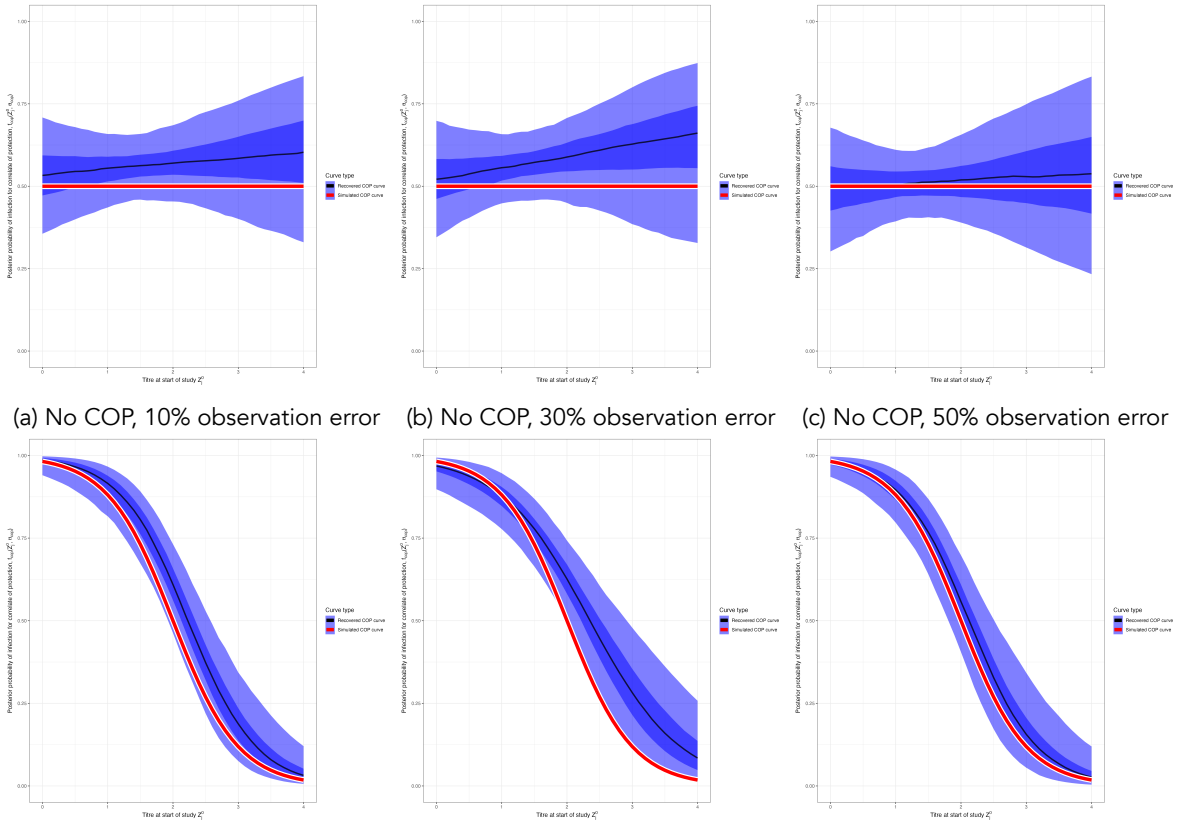


(d) COP, 10% observation error (e) COP, 30% observation error (f) COP, 50% observation error

Figure 1: 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%)

1.5.2 Correlate of protection

We next assess the ability of **Algorithm ??** 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 ??**. Here, it is clear, although we have a pre-defined function for the correlate of protection, when we simulate the data, the COP cruve is not perfectly recovered, and thus, the posterior distribution of our inference method can, at best, recover the inferred COP curve.



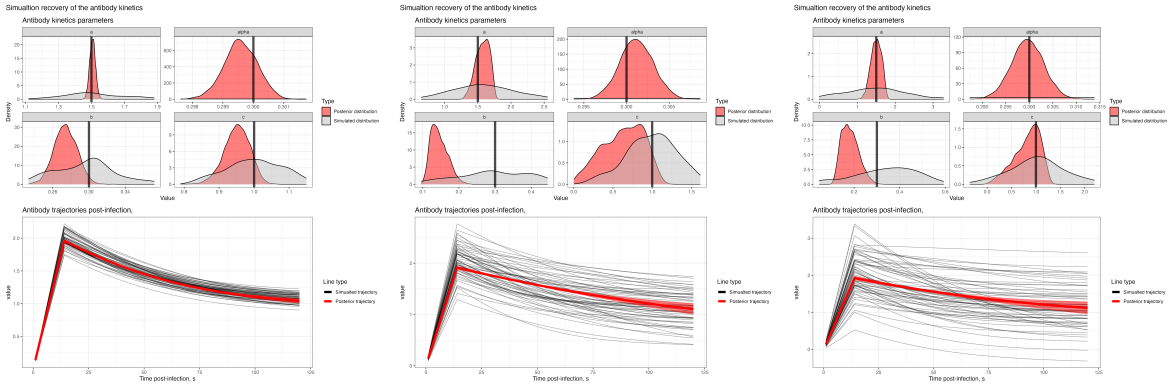
(a) No COP, 10% observation error (b) No COP, 30% observation error (c) No COP, 50% observation error

(d) COP, 10% observation error (e) COP, 30% observation error (f) COP, 50% observation error

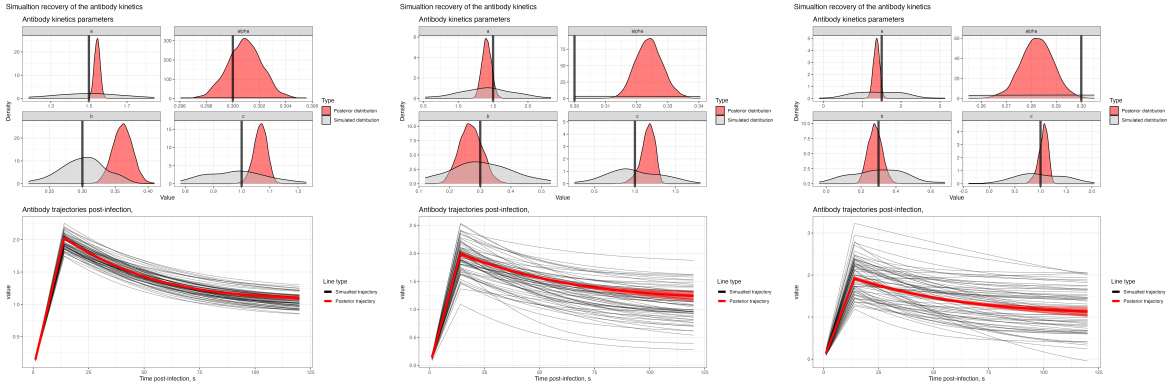
Figure 2: Simulation recovery of the COP function, with posterior samples plot $f_{cop}(x, \hat{\theta}_{cop})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).

1.5.3 Antibody kinetics

Algorithm ?? also succesfully recovers the simualted antibody kinetics. Let us plot $f^1_{ab}(s, \hat{a}, \hat{b}, \hat{c})$, 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 ??**).



(a) No COP, 10% observation error (b) No COP, 30% observation error (c) No COP, 50% observation error



(d) COP, 10% observation error (e) COP, 30% observation error (f) COP, 50% observation error

Figure 3: Simulation recovery of the antibody kinetics function with posterior samples plot $f^1_{ab}(s, \hat{a}, \hat{b}, \hat{c})$. We have two different COP models (top: No COP, bottom: logistic COP) and three different levels of antibody kinetics variability (10%, 30%, 50%).