

Efficient likelihood approximation via Gaussian processes:  
with an application to an existing *Plasmodium vivax*  
malaria model

The University of Melbourne

Jacob Cumming

May 2024

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>I</b>	<b>Literature Review</b>	<b>3</b>
<b>2</b>	<b>Epidemiological Modelling</b>	<b>5</b>
2.1	Deterministic Ordinary Differential Equation Models . . . . .	6
2.2	Stochastic models . . . . .	8
<b>3</b>	<b>Malaria and Malaria Models</b>	<b>13</b>
3.1	Mathematical Modelling of Malaria . . . . .	14
<b>4</b>	<b>Parameter Inference</b>	<b>21</b>
4.1	Motivation . . . . .	21
4.2	Frequentist Parameter Estimation . . . . .	21
4.3	Bayesian Parameter Estimation . . . . .	25
<b>5</b>	<b>Gaussian Processes and Synthetic Likelihoods</b>	<b>35</b>
5.1	Gaussian Processes . . . . .	35
5.2	Gaussian Process Regression . . . . .	40
5.3	Model Selection . . . . .	44
5.4	Bayesian Acquisition Functions . . . . .	45
<b>II</b>	<b>Calibrating Parameters for a <i>P. vivax</i> Model</b>	<b>47</b>
<b>6</b>	<b>Methods</b>	<b>49</b>
6.1	Creation of Synthetic Data . . . . .	49
6.2	Model Simulations and Discrepancy Function . . . . .	50
6.3	Gaussian Process and Initialisation . . . . .	51
6.4	Bayesian Acquisition and Parameter Updates . . . . .	53
<b>7</b>	<b>Results</b>	<b>55</b>
7.1	Validation . . . . .	55
7.2	Parameter estimation . . . . .	55
<b>8</b>	<b>Discussion</b>	<b>61</b>

<b>9 Conclusion</b>	<b>65</b>
<b>Bibliography</b>	<b>65</b>
<b>A Additional Theorems and Proofs</b>	<b>71</b>
<b>B Additional Results</b>	<b>73</b>

# List of Tables

6.1	The parameters used to simulate a <i>P. vivax</i> outbreak using the model described by Champagne et al. 2022 . . . . .	50
6.2	Observed synthetic data $\mathbf{y}^{\text{obs}} := \{\iota_{\text{obs}}, \pi_{\text{obs}}, i_{\text{obs}}, p_{\text{obs}}\}$ from the simulation in Figure 6.1. . . . .	50
6.3	Conservative upper bounds for parameters to be calibrated. Values were informed by Champagne et al. 2022; White et al. 2016. All lower bounds were zero. . . . .	51
6.4	Hyperparameters used in training $d_{\mathcal{GP}}(\boldsymbol{\theta})$ . . . . .	52
7.1	Final Gaussian process hyperparameters . . . . .	55
7.2	Estimates of our model parameters. The maximum likelihood estimate (MLE) of the true parameters using $\hat{L}$ . The maximum slice estimate was the one-dimensional maximum likelihood estimate where all other parameters are held constant at the true value. . . . .	56



# List of Figures

2.1	Some simple model schematics with varying numbers of compartments: $S$ (susceptible), $E$ (exposed), $I$ (infectious) and $R$ (recovered). The force of infection $\lambda_t$ is usually a function of $I_t$ , depicted by the dashed red lines. $\mu$ and $\nu$ are natural birth and death rates, respectively. $\gamma$ is the rate of progression out of the infectious state. In each of these models, the physical interpretation differs slightly. In the $SIS$ and $SEIR$ models, it is the rate at which individuals move from infectious to susceptible again or into lifelong immunity, whereas, in the $SI$ with demography model, it can be interpreted as the increase in the rate of death attributable to disease-induced mortality. $\sigma$ is the rate of progression from a state of latent infection to becoming infectious. . . . .	6
2.2	Solutions to the ODEs describing the models depicted in Figure 2.1. The initial infectious population was $I_0 = 10$ , with $S_0 = 990$ . In the $SEIR$ model, $E_0 = R_0 = 0$ . For all models $\beta = 0.4$ . For the $SIS$ and $SI$ model with demography $\gamma = 1/4$ . For the $SI$ model with demography $\mu = 0.012$ , and $\nu = 0.0012$ . For the $SEIR$ model, $\gamma = 1/90$ , and $\sigma = 1/2$ . . . . .	7
2.3	Exact stochastic simulations of the 3 different models using Algorithm 1. The parameters used were identical to those in Figure 2.2 . . . . .	10
3.1	The <i>P. vivax</i> (malaria) lifecycle. <i>P. falciparum</i> does not have a dormant liver hypnozoite stage. Created with BioRender.com. . . . .	14
3.2	A simple Ross-Macdonald malaria model schematic, as described by Aron and May 1982. $S_H$ and $I_H$ are the number of susceptible and infected humans respectively, and $S_M$ and $I_M$ are the number of susceptible and infected mosquitos. The rate of human infection ( $\lambda_H$ ) is dependant on $I_M$ , and the rate of human infection ( $\lambda_M$ ) is dependant on $I_H$ . . . . .	15
3.3	Diagram for <i>P. vivax</i> model in a tropical setting described by White et al. 2016. $S$ and $I$ are the number of susceptible and infected humans and mosquitos (denoted by subscript M). $\lambda_H = mabI_M$ and $\lambda_M = ac(I_0 + I_L)$ . . . . .	16
3.4	Diagram for <i>P. vivax</i> model described by Champagne et al. 2022. $I_{\text{total}} = I_0 + I_L$ . Since the mosquito dynamics have been removed, $\lambda$ now not has no dependencies on the number of infectious mosquitos. . . . .	18

- 4.1 Two linear models of the form  $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$  fit given the set of observations  $\{(1, 2), (2, 4), (3, 4)\}$  using the method of least squares and maximum likelihood under the assumption that the data are independent realisations from a Poisson distribution with mean  $y_i(\boldsymbol{\theta})$ . The least squares estimates were  $\theta_0^{\text{LSE}} = 4/3$  and  $\theta_1^{\text{LSE}} = 1$ . The maximum likelihood estimates were  $\hat{\theta}_0 \approx 1.329$  and  $\hat{\theta}_1 \approx 0.751$ . . . . . 22
- 4.2 An *SEIR* model fit to some observed prevalence data taken every two weeks over a 14 week period, generated as  $I_t/N$  from the *SEIR* simulation in Figure 2.3. All parameters were considered known except for  $\beta$ . The least squares estimate (LSE)  $\beta^{\text{LSE}} = 0.3516$  was found by solving the model ODEs and numerically minimising the square differences between observed prevalences and the ODE prevalences (as proportions). Similarly the maximum likelihood estimate  $\hat{\beta} = 0.3493$  was found by assuming the prevalence (times 1000) was binomially distributed from 1000 samples with the probability of being infectious equal to  $\frac{I_t}{N}$ . . . . . 25
- 4.3 Samples of  $X$  from the unnormalised density  $g(x) = (x - 1)^2$  with  $x \in (0, 2)$  using the rejection sampler.  $X^* \sim \text{Unif}(0, 2)$  and  $M = 1$ . Green dots are samples from from  $X$ . Of 500 samples of  $X^*$ , 157 were accepted as samples of  $X$ . . . . . 27
- 4.4 A simple time homogeneous Markov chain, with two states. It is characterised by the transition kernel  $K(1, 1) = \Pr(X_{i+1} = 1|X_i = 1) = 0.7$ ,  $K(1, 2) = \Pr(X_{i+1} = 2|X_i = 1) = 0.3$ ,  $K(2, 1) = \Pr(X_{i+1} = 1|X_i = 2) = 0.4$ , and  $K(2, 2) = \Pr(X_{i+1} = 2|X_i = 2) = 0.6$ . The stationary distribution is  $\pi(1) = 4/7$  and  $\pi(2) = 3/7$ . . . . . 27
- 4.5 Samples from the posterior distribution of  $p$  using the Metropolis-Hastings algorithm.  $p$  was assumed to have a uniform prior between 0 and 1, with  $y^{\text{obs}} = 6$ , generated from  $\text{Binom}(10, p)$ , The choice of proposal distribution did not impact the final estimate of  $\Pr(p|y^{\text{obs}})$ . . . . . 29
- 4.6 Given a daily incidence of  $y^{\text{obs}} = 26$  at day 30 of an *SIS* epidemic, with unknown  $\beta$ , we use Metropolis-Hasting to sample from  $\Pr(\beta|y^{\text{obs}})$ .  $\gamma = 1/4$  was assumed to be correct, and we compared the assumption  $y^{\text{obs}} \sim \text{Binom}(\lfloor S_{30} \rfloor, \beta I_{30}/N)$ , to the assumption  $y^{\text{obs}} \sim \text{Pois}(\frac{\beta I_{30} S_{30}}{N})$  where  $I_{30}, S_{30}$  are the ODE solutions to Equations 2.1 and 2.2. We assumed the prior distribution  $\beta \sim \text{Gamma}(2, 6)$ , where  $\mathbb{E}(\beta) = 1/3$ . Our proposal density was  $N(\beta^*, 1/10)$ , where  $\beta^*$  was the previous sample. . . . . 31
- 4.7 2000 posterior samples from  $\Pr(\beta, \gamma|\mathbf{y}^{\text{obs}})$ , where  $\beta|\gamma, \mathbf{y}^{\text{obs}} \sim \text{Gamma}(9, 4/\gamma + 4 + 8\gamma)$  and  $\gamma|\beta, \mathbf{y}^{\text{obs}} \sim \text{InvGamma}(12, 12\beta)$ . The samples were obtained using a Gibbs sampler. The red points are the first 15 samples using the Gibbs sampler. . . . . 32
- 5.1 Ten sample realisations from 4 different kernels, with one realisation bolded. Samples for each kernel were generated from the same seed and the hyperparameters  $\ell$ , and  $\sigma_k$  were set to 1. . . . . 37
- 5.2 Ten realisations of zero mean Gaussian processes with the squared exponential kernel, varying the length and amplitude parameters. The samples were generated using the same seed . . . . . 39
- 5.3 Sequence of Gaussian process regressions on the target function (black)  $f(x) = x(x - 1)(x + 1)$ , after 1, 2, 4, and 8 observations in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was zero mean and had a squared exponential kernel. The hyperparameters were fixed with  $\ell = 2.7$  and  $\sigma_k^2 = 1.1$  . . . . . 42

5.4	Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$ , after 2, 4, 8, and 16 observations of $f(x_i) + \varepsilon_i$ , where $\varepsilon_i$ is i.i.d. $\text{MVN}(0, \sigma_o^2)$ with $\sigma_o^2 = 0.01$ in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was 0 mean and had a squared exponential kernel. The hyperparameters were fixed with $\ell = 2.7$ and $\sigma_k^2 = 1.1$ . . .	43
6.1	A Doob-Gillespie Simulation of the model described by Champagne et al. 2022 with $\alpha = 0.4$ , $\beta = 0.4$ , $\gamma_L = 1/223$ , $\lambda = 0.04$ , $f = 1/72$ , $r = 1/60$ , and $\delta = 0$ . The population was 10000, with 100 initial infections (both blood and liver stage $I_L$ ). . . . .	49
7.1	$\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$ violin plot . . . . .	56
7.2	Hyperparameter training . . . . .	56
7.3	Finding $\arg \min_{\boldsymbol{\theta}} \mathcal{A}_{\text{EI}}(\boldsymbol{\theta})$ . . . . .	56
7.4	The left column of figures is the Gaussian process after initialisation $d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta})$ . The black line is $\mathbb{E}(d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta}))$ , and the red lines are multiple realisations of $d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta})$ . The right column of figures is after 400 sampling iterations, with the black line being $\mathbb{E}(d_{\mathcal{GP}}^{(400)}(\boldsymbol{\theta}))$ . The blue dots are realisations of $\ln \mathcal{D}(\boldsymbol{\theta})$ . $d_{\mathcal{GP}}$ has not been trained on these realisations. The parameters are varied univariately, with all other parameters fixed at the true parameters. . . . .	57
7.5	Gaussian process approximations of the treatment parameters, as with Figure 7.4 . . . . .	58
7.6	Final univariate synthetic likelihoods $\hat{L}(\boldsymbol{\theta})$ after 400 sampling iterations. All values not shown were fixed at the true parameters. . . . .	59
B.1	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $\alpha$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	74
B.2	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $\beta$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	75
B.3	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $\gamma_L$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	76
B.4	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $\lambda$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	77
B.5	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $f$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	78
B.6	$d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , for $t = 0, 100, 200, 300$ , and 400. Only $r$ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$ . Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$ . . . . .	79



# Chapter 1

## Introduction

Malaria is an infectious disease that poses a significant global health challenge. In 2022, the World Health Organisation estimated that malaria caused over 600,000 deaths, most of which were in children under five (World Health Organization 2022). The majority of malaria-related deaths are attributable to the *Plasmodium falciparum* species of the parasite. However, the literature is increasingly recognising that the amount of death and severe disease attributable to the *Plasmodium vivax* species is likely traditionally underestimated. The malaria lifecycle is complicated, needing both a vertebrate and mosquito host. In addition, *P. vivax* has a dormant stage that causes relapses.

Various countries with endemic malaria have undertaken significant efforts to reduce the burden of or eradicate malaria. Mathematical disease models are increasingly aiding these efforts by helping to understand the spread of the disease and estimate the effectiveness of possible interventions. Malaria models are complex due to the many staged lifecycle, and in particular, *P. vivax* models need to consider relapses, further complicating the model.

To simulate different scenarios using models, modellers must first calibrate the model parameters to approximate observed disease dynamics, measured by observed data such as case counts or prevalence surveys. Standard techniques to calibrate parameters include maximum likelihood estimation and sampling from a posterior distribution, which both require a likelihood function. As models become increasingly complicated, analytic forms for the likelihood may not exist, or calculating the likelihood may be very computationally burdensome. Some researchers calibrate compartmental models by relying on their deterministic counterparts, which is questionable, as the deterministic model sometimes behaves differently from the stochastic model.

Modern likelihood-free techniques, such as approximate Bayesian computation, have been designed to facilitate a principled method of parameter calibration. Various forms of approximate Bayesian computation have been widely adopted, however forms require large numbers of model runs, which may be unfeasible or undesirable.

Drawing on concepts from approximate Bayesian computation, we aimed to improve parameter calibration in malaria models, addressing a recognised need for improvement. We do this by training a Gaussian process to predict how close a model run will be to observed data. The Gaussian process approximation can then be used to extract a synthetic likelihood which approximates the true unknown likelihood. This allows for the use of frequentist and Bayesian parameter inference.

The thesis follows the following outline: The literature review in Part I discusses fundamental concepts of epidemiological modelling, covering deterministic ordinary differential equation mod-

els, stochastic models, and their simulation methods. It then examines malaria and reviews various malaria models. Part I further surveys parameter inference techniques, comparing frequentist and Bayesian approaches, and ends with discussing likelihood-free techniques, particularly approximate Bayesian computation. Approximate Bayesian computation then motivates using Gaussian processes to develop a synthetic likelihood, which can be used in place of the unknown true likelihood to use the traditional parameter calibration techniques in complex models. Finally, Part I discusses using Bayesian acquisition functions to train the Gaussian process efficiently.

Part II applies and extends this methodology by calibrating parameters specific to a *P. vivax* model using synthetic observed data. The results and discussion section validates the method and demonstrates that the parameters that produced the observed data are recoverable. Finally, the thesis concludes with a discussion of the findings and outlines avenues for future research.

**Part I**

**Literature Review**



## Chapter 2

# Epidemiological Modelling

Researchers have developed compartmental epidemiological models to study the behaviour and characteristics of disease spread and eradication. Compartmental models simplify the dynamics of a disease down to a mathematically representable form. By determining the parameters within a model, it is possible to facilitate an understanding of how the modelled disease spreads and assess the effectiveness of differing disease interventions (such as treatments or vaccinations) without large long-term trials. Models can also simulate various scenarios, such as increases or decreases in viral transmission.

Simple compartmental disease models assume individuals can only be in one of a finite number of states (which are called compartments). These compartments usually correspond to a state of disease. A standard compartmental model may include:

- $S$  - Susceptible: at risk of contracting the disease
- $E$  - Exposed: contracted the disease but not infecting other individuals
- $I$  - Infectious (also called Infected): at risk of infecting others with the disease
- $R$  - Recovered: no longer infectious and incapable of being reinfected.

The number of people in each compartment at time  $t$  is a (possibly non-deterministic) function of time  $t$ , which we indicate as a subscript  $t$  (e.g.,  $S_t$  is the number of susceptibles at time  $t$ ). Models are typically described by the compartments they contain. The  $SIS$  model depicted in Figure 2.1a consists of a susceptible compartment  $S$ , and an infectious compartment  $I$ . In this model, individuals who recover from infection are immediately susceptible to reinfection, as is the case with most sexually transmitted diseases (Keeling and Rohani 2008, p. 56). Movement between the two compartments determined by the force of infection  $\lambda_t$  and rate of recovery  $\gamma$ .

The  $SI$  with demography model depicted in Figure 2.1b is used to model diseases that infect the individual until the time of death, such as bovine spongiform encephalopathy (BSE, commonly known as mad cow disease) (Hagenaars, Donnelly, and Ferguson 2006). In addition to the rates in the  $SIS$  model,  $\mu$  and  $\nu$  are birth and death rates respectively.  $\gamma$  becomes the rate of disease induced mortality.

The  $SEIR$  model depicted in Figure 2.1c includes the exposed compartment  $E$  and recovered compartment  $R$ . The additional parameter  $\sigma$  is the rate at which individuals who are exposed become infectious. Childhood diseases such as varicella (chickenpox) which give lifetime immunity after infection, can be modelled using an  $SEIR$  model (see Figure ), particularly when modelling

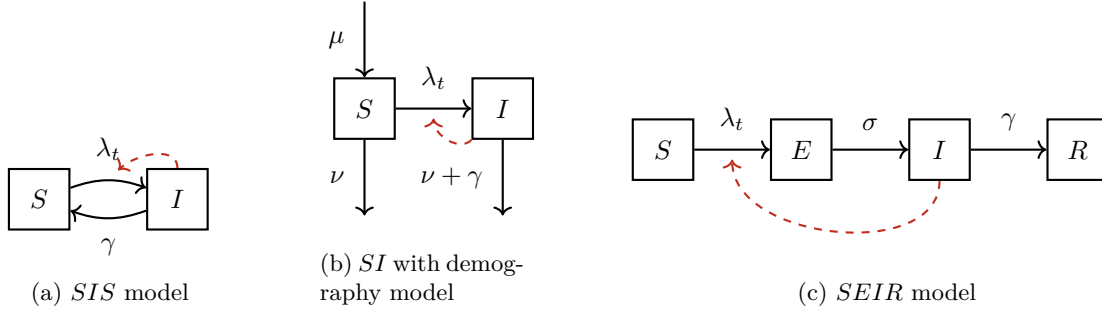


Figure 2.1: Some simple model schematics with varying numbers of compartments:  $S$  (susceptible),  $E$  (exposed),  $I$  (infectious) and  $R$  (recovered). The force of infection  $\lambda_t$  is usually a function of  $I_t$ , depicted by the dashed red lines.  $\mu$  and  $\nu$  are natural birth and death rates, respectively.  $\gamma$  is the rate of progression out of the infectious state. In each of these models, the physical interpretation differs slightly. In the *SIS* and *SEIR* models, it is the rate at which individuals move from infectious to susceptible again or into lifelong immunity, whereas, in the *SI* with demography model, it can be interpreted as the increase in the rate of death attributable to disease-induced mortality.  $\sigma$  is the rate of progression from a state of latent infection to becoming infectious.

a local outbreak setting. For example, an *SEIR* model was used in Zha et al. 2020 for a school outbreak of varicella.

Omitting demography is usually appropriate when disease-induced mortality is low, or the timescale of interest is short compared to the population's lifespan. The number of compartments in a model can be increased as required, to incorporate states such as vaccinated, or quarantined that change the disease dynamics. For example, the Covid-19 model described in Acuña-Zegarra et al. 2021 includes a vaccinated compartment, and two infectious compartments for symptomatic and asymptomatic individuals. By convention,  $N_t$  (often simply  $N$  in models with a closed population) is the total number of individuals in the model - the sum of all compartments.

## 2.1 Deterministic Ordinary Differential Equation Models

Infectious diseases are often simulated as deterministic ordinary differential equations (ODEs). Let the force of infection  $\lambda_t$  be proportional to the number of people in  $I$ , such that  $\lambda_t := \beta \frac{I_t}{N_t}$ .  $\beta$  can be interpreted as the average number of people that an individual interacts with per day in a way such that disease would be spread in that interaction per unit of time  $t$ . Since  $\frac{I_t}{N_t}$  is the probability that a randomly selected individual is infectious,  $\beta \frac{I_t}{N_t}$  can be interpreted as the average number of people that a person interacts with each day who are infectious multiplied by the probability that an infection occurs during that contact. In different diseases  $\beta$  varies dramatically.  $\beta$  is low for diseases that need prolonged exposure or sexual contact to transmit, whereas  $\beta$  is very high for diseases that are very highly transmittable such as measles. Implicitly there is an assumption of complete uniformly random mixing of people. This is often a poor assumption, since in reality people see people regularly with varying probabilities, and do not mix randomly with the people in a population. This means the infection status of a group of friends or a household is likely to be highly correlated. One solution would be to include contact networks in the model (such as Kerr et al. 2021), however we do not consider these models in this thesis. For this thesis we assume that  $\beta$  is frequency dependent as opposed to density dependent, that is, a person interacts with the same number of individuals regardless of population size. A density dependent contact rate

assumes that the number of individuals a person interacts with grows proportional to population size. In the density dependent case the force of infection has the form  $\lambda_t := \beta I_t$ .

The ODEs that govern the *SIS* model, are

$$\frac{dS_t}{dt} = -\lambda S_t + \gamma I_t = -\beta \frac{I_t}{N} S_t + \gamma I_t \quad (2.1)$$

$$\frac{dI_t}{dt} = \lambda S_t - \gamma I_t = \beta \frac{I_t}{N} S_t - \gamma I_t. \quad (2.2)$$

Given a set of initial conditions  $S_0$  and  $I_0$ , Equation 2.1 (or 2.2) fully describes the model.

The system of ODEs that describe the *SI* with demography model is

$$\frac{dS_t}{dt} = \mu N_t - \lambda_t S_t - \nu I_t = \mu N_t - \beta \frac{I_t}{N_t} S_t - \nu I_t \quad (2.3)$$

$$\frac{dI_t}{dt} = \lambda_t S_t - (\gamma + \nu) I_t = \beta \frac{I_t}{N_t} S_t - (\gamma + \nu) I_t. \quad (2.4)$$

Unlike for the *SIS* and *SEIR* models,  $N_t$  is not constant in this model.

Finally, the system of ODEs that describe the *SEIR* model is

$$\frac{dS_t}{dt} = -\lambda_t S_t - \nu I_t = -\beta \frac{I_t}{N} S_t + \gamma I_t \quad (2.5)$$

$$\frac{dE_t}{dt} = \lambda_t S_t - \omega E_t = \beta \frac{I_t}{N} S_t - \omega I_t \quad (2.6)$$

$$\frac{dI_t}{dt} = \omega E_t - \gamma I_t \quad (2.7)$$

$$\frac{dR_t}{dt} = \gamma I_t \quad (2.8)$$

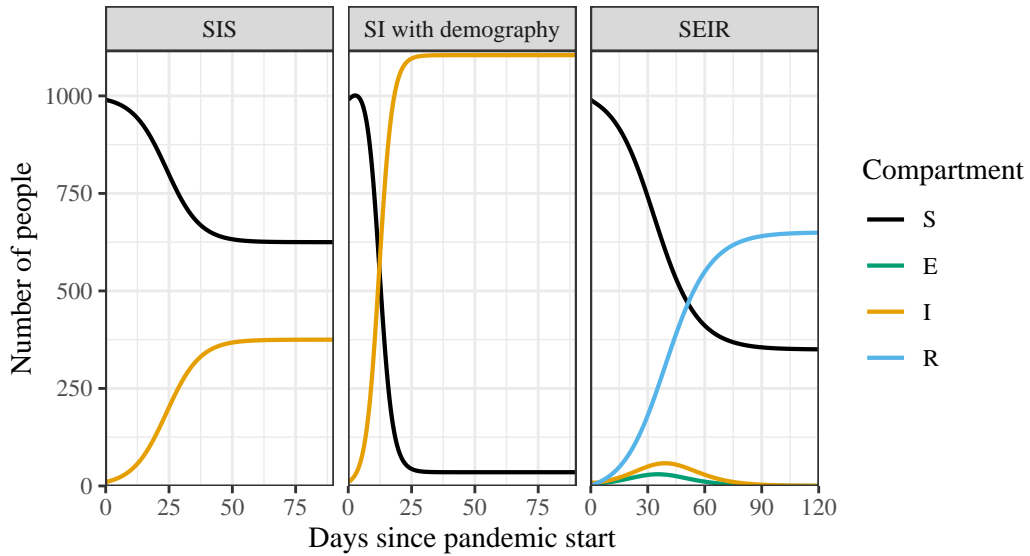


Figure 2.2: Solutions to the ODEs describing the models depicted in Figure 2.1. The initial infectious population was  $I_0 = 10$ , with  $S_0 = 990$ . In the *SEIR* model,  $E_0 = R_0 = 0$ . For all models  $\beta = 0.4$ . For the *SIS* and *SI* model with demography  $\gamma = 1/4$ . For the *SI* model with demography  $\mu = 0.012$ , and  $\nu = 0.0012$ . For the *SEIR* model,  $\gamma = 1/90$ , and  $\sigma = 1/2$ .

After specifying the the initial conditions  $S_0, I_0$  etc. the ODEs can be numerically solved. Figure 2.2 shows solutions to the three models introduced so far for an initial population  $N_0 = 1000$ , with an initial infectious population  $I_0 = 10$ . Under the models without demography,  $N_t$  remains constant. The disease is eradicated for the  $SEIR$  model after a small spike in cases. The disease reaches a steady state equilibrium for the  $SIS$  and  $SI$  with demography model, with a much higher number of infected individuals in the  $SI$  with demography model, since once an individual acquires the disease they have it until they die.

## 2.2 Stochastic models

### Motivating the form of the stochastic model

Deterministic ODE models can appropriate to study infectious diseases when the disease is near equilibrium and the numbers in each compartment are large. At the start of an epidemic, when the number of infected individuals is small the behaviour of the epidemic may vary significantly. When case numbers are small, there is a non zero probability that the disease may die out, but under the right conditions such as a large gathering of people in a small space, the outbreak may become a pandemic. In these cases, a deterministic model is inadequate in emulating the real world behaviour. For this reason we consider models of the disease which are stochastic in nature. A natural stochastic analogue for a deterministic model can be constructed through consideration of Poisson point processes and their properties.

**Definition 2.1** (Poisson Point Process).  $\{\mathcal{N}(t)\}_{t \geq 0}$  is a (stationary) Poisson point process with intensity  $\kappa$  if

1.  $\mathcal{N}(0) = 0$
2.  $\mathcal{N}(t_1) - \mathcal{N}(t_0), \mathcal{N}(t_2) - \mathcal{N}(t_1), \dots, \mathcal{N}(t_n) - \mathcal{N}(t_{n-1})$  are independent for  $0 \leq t_0 < t_1 < \dots < t_{n-1} < t_n$
3.  $\mathcal{N}(t_2) - \mathcal{N}(t_1) \sim \text{Pois}(\kappa(t_2 - t_1)), 0 \leq t_1 < t_2$ .

To demonstrate how Poisson point processes aid us in creating a natural analogue of the deterministic model, we construct two Poisson processes that when combined have the the same instantaneous average behaviour as the deterministic ODE model.

Consider the deterministic  $SIS$  model described by Equations 2.1 and 2.2 at time  $t^*$ . The instantaneous rate at which  $S_t$  is decreasing is  $\beta \frac{I_{t^*}}{N} S_{t^*}$ . In other words, at time  $t^*$ , an individual leaves the susceptible compartment every  $\beta \frac{I_{t^*}}{N} S_{t^*}$  units of time.

Now consider a Poisson point process  $\{\mathcal{N}_1(t - t^*)\}_{t \geq t^*}$  with intensity  $\beta \frac{I_{t^*}}{N} S_{t^*}$  corresponding to the count of the number of individuals who have left  $S$  and entered  $I$   $t$  units of time since  $t^*$ . The average rate at which an individual leaves  $S$  is then

$$\begin{aligned} \frac{d\mathbb{E}(\mathcal{N}_1(t^*))}{dt^*} &= \lim_{\delta \rightarrow 0} \frac{\mathbb{E}(\mathcal{N}(t^* + \delta) - \mathcal{N}(t^*))}{\delta} \\ &\quad \text{(Since the mean of a Poisson random variable is its intensity)} \\ &= \frac{\beta \frac{I_{t^*}}{N} S_{t^*} (t^* + \delta - t^*) \beta \frac{I_{t^*}}{N} S_{t^*}}{\delta} \\ &= \beta \frac{I_{t^*}}{N} S_{t^*}, \end{aligned}$$



the same rate as the ODE model.

Under the same deterministic ODE formulation of the *SIS* model, the instantaneous rate into *S* at time  $t^*$  is  $\gamma I_{t^*}$ . Also as above we construct a Poisson point process  $\{\mathcal{N}_2(t-t^*)\}_{t \geq t^*}$  with mean rate  $\gamma I_{t^*}$  describing the number of recoveries from *I* to *S*, since

$$\frac{d\mathbb{E}(\mathcal{N}_2(t^*))}{dt^*} = \gamma I_{t^*}.$$

Combining the two processes, we can see that the rate of change in the average number of people in *S* is

$$\frac{d\mathbb{E}(\mathcal{N}_2(t^*) - \mathcal{N}_1(t^*))}{dt^*} = \frac{d\mathbb{E}(\mathcal{N}_2(t^*)) - d\mathbb{E}(\mathcal{N}_1(t^*))}{dt^*} = -\beta \frac{I_t}{N} S_t + \gamma I = \frac{dS_t}{dt}.$$

For an arbitrary number of compartments and transitions, we can model each transition as a Poisson point process.

Therefore we construct a stochastic analogue to the ODEs in the following way. Let the stochastic model be a random vector  $\{\mathbf{C}_t\}_{t \geq 0} = \{C_1(t), C_2(t), \dots, C_n(t)\}_{t \geq 0}$  where  $C_i : \mathbb{R} \rightarrow \mathbb{N} \cup \{0\}$ , is the number of people in compartment  $C_i$ , and for any fixed  $t$ ,  $\{C_1(t), C_2(t), \dots, C_n(t)\}$  is a random variable describing the state of the model. For example the *SI* with demography model can be represented as  $\{\mathbf{C}_t\}_{t \geq 0} := \{S_t, I_t\}_{t \geq 0}$ .

If the model is in state  $\{S_t, I_t\}$ , after the next transition at time  $t^*$ , the next possible states are:

1.  $\{S_{t^*}, I_{t^*}\} = \{S_t + 1, I_t\}$
2.  $\{S_{t^*}, I_{t^*}\} = \{S_t - 1, I_t\}$
3.  $\{S_{t^*}, I_{t^*}\} = \{S_t - 1, I_t + 1\}$
4.  $\{S_{t^*}, I_{t^*}\} = \{S_t, I_t - 1\}$

Each of these transitions behave like Poisson processes at time  $t$ . We think of these processes in the following way:

- $\{\mathcal{E}_1(t^*)\}_{t^* \geq 0}$  : the number of births into *S* after time  $t$  with intensity  $\mu N_t$
- $\{\mathcal{E}_2(t^*)\}_{t^* \geq 0}$  : the number of deaths in *S* after time  $t$  with intensity  $\nu S_t$
- $\{\mathcal{E}_3(t^*)\}_{t^* \geq 0}$  : the number of infections after time  $t$  with intensity  $\beta \frac{I_t}{N_t} S_t$
- $\{\mathcal{E}_4(t^*)\}_{t^* \geq 0}$  : the number of deaths from *I* after time  $t$  with intensity  $(\nu + \gamma) I_t$ .

Since the sum of two Poisson point processes is a Poisson point process (see Theorem A.1) and the time between events in a Poisson process is exponentially distributed (see Theorem A.2).

$$\{\mathcal{E}(t)\}_{t \geq 0} := \{\mathcal{E}_1(t) + \mathcal{E}_2(t) + \mathcal{E}_3(t) + \mathcal{E}_4(t)\}_{t \geq 0}$$

is a Poisson point process with intensity

$$\mu N_{t^*} + \nu S_{t^*} + \beta \frac{I_{t^*}}{N_{t^*}} S_{t^*} + (\nu + \gamma) I_{t^*},$$

and the time until the next transition is exponentially distributed

$$\text{Exp}(\mu N_{t^*} + \nu S_{t^*} + \beta \frac{I_{t^*}}{N_{t^*}} S_{t^*} + (\nu + \gamma) I_{t^*}).$$

Therefore simulating the time to the next transition given the current number of individuals in each compartment just involves simulating from an exponential distribution. Theorem A.3 states that given that a transition occurred, the probability it was due to the process  $\mathcal{E}_i$ , is proportional to its intensity. As soon as the transition occurs, all the intensities for each Poisson process will be updated, since the intensities are dependent on the number of individuals in each compartment.

## Doob-Gillespie Algorithm

---

### Algorithm 1 The Doob-Gillespie Algorithm

---

Initialise time  $t \leftarrow 0$  and initial state of the model  $\mathbf{C}(0) := \{C_1(0), C_2(0), \dots, C_n(0)\}$

**while** termination condition not met **do**

    Calculate intensities  $\kappa_i$  for all possible events  $\mathcal{E}_i$

    Calculate total intensity  $\kappa = \sum_i \kappa_i$

    Generate  $\Delta t \sim \text{Exp}(\kappa)$

    Choose event  $\mathcal{E}_i$  with probability  $\frac{\kappa_i}{\kappa}$

    Update time  $t \leftarrow t + \Delta t$

    Update state of  $\mathbf{C}(t + \delta t) \leftarrow \mathbf{C}(t) + \text{change in state due to event } \mathcal{E}_i$

**end while**

---

All of this leads naturally to a common method of simulating the stochastic model through Algorithm 1. The Doob-Gillespie algorithm exploits the local behaviour of the compartments as described, sampling an exponential random variable, choosing an event with probabilities proportional to their intensities, updating the model, and updating the intensities given a set of starting conditions (Gillespie 1977).

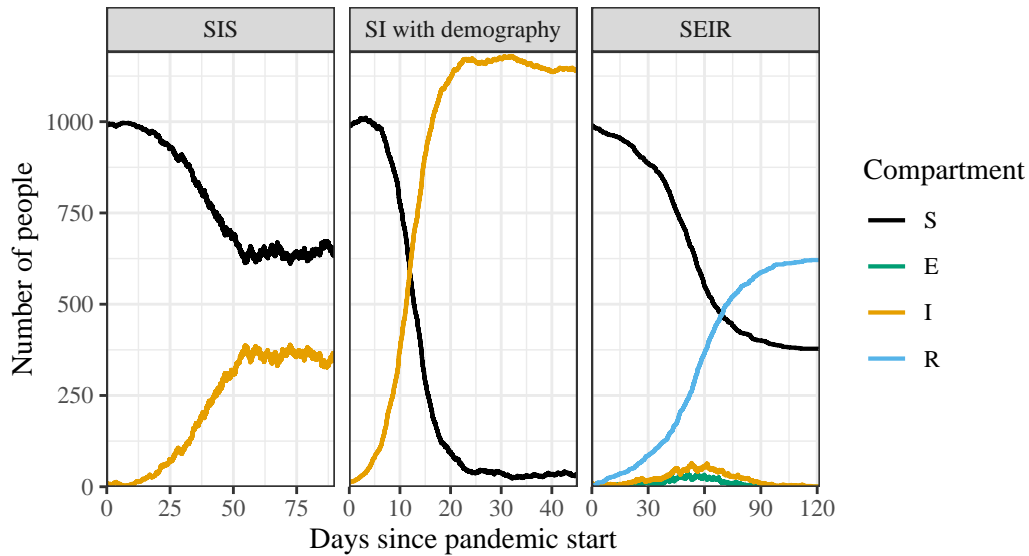


Figure 2.3: Exact stochastic simulations of the 3 different models using Algorithm 1. The parameters used were identical to those in Figure 2.2

Figure 2.3 demonstrates a Doob-Gillespie simulations for the models discussed this chapter.

Compared to the deterministic ODE simulations in Figure 2.2, the stochastic simulations still has variation even near the equilibria.

### $\tau$ -leaping

Because the Doob-Gillespie algorithm is an exact simulation, when the intensities become large or many people are getting infected, the time step  $\tau$  can become increasingly small. This can result in simulations taking a long time.

$\tau$ -leaping - Algorithm 2 - further exploits the local Poisson point process like behaviour of epidemiological models (Gillespie 2001). Consider the SIS model, when  $S_t = I_t = 10000$ . Events happen at a very high rate, meaning the  $\Delta t$  found in each step of the Doob-Gillespie algorithm will be very small, but the rates also change a negligible amount after each event (compare  $\gamma \times 10000$  to  $\gamma \times 10001$  or  $\gamma \times 9999$ ). Therefore we can approximate the number of events in a short time period  $\tau$  as a Poisson point process with the total intensity  $\kappa = \sum_i \kappa_i$  at time  $t$ , with the probability of any one event having the same probability as above of  $\frac{\kappa_i}{\kappa}$ . Therefore we have the following algorithm.

---

#### **Algorithm 2** $\tau$ -Leaping Algorithm Gillespie 2001

---

Initialise time  $t \leftarrow 0$  and initial state of the model  $\mathbf{C}(0) := \{C_1(0), C_2(0), \dots, C_n(0)\}$

**while** termination condition not met **do**

    Calculate intensities  $\kappa_i$  for all possible events  $\mathcal{E}_i$

    Calculate total intensity  $\kappa = \sum_i \kappa_i$

    Choose a suitable time step  $\tau$  (this can be deterministic or adaptive)

    Calculate Poisson random variable  $X \sim \text{Poisson}(\kappa\tau)$

**for**  $i$  in 1 to  $X$  **do**

        Choose event  $\mathcal{E}_i$  with probability  $\frac{\kappa_i}{\kappa}$

        Update state of  $\mathbf{C}(t + \tau) \leftarrow \mathbf{C}(t) + \text{change in state due to event } \mathcal{E}_i$

**end for**

    Update time  $t \leftarrow t + \tau$

**end while**

---



## Chapter 3

# Malaria and Malaria Models

Malaria as an infectious disease has been the focus of mathematical modelling efforts for over a century (Smith et al. 2012). Despite ongoing eradication efforts, in 2022 malaria killed over 600,000 people, with over 75% of deaths occurring in children under 5 years old (World Health Organization 2022). Six species of the parasite are able to infect humans (Milner 2018). Although *Plasmodium falciparum* is responsible for around 90% of total human malaria deaths, outside of Africa *Plasmodium vivax* is the leading cause of malaria infection (Zekar and Sharman 2023; Adams and Mueller 2017). It is likely that death and severe disease attributable to *P. vivax* has been traditionally underestimated. In view of recent evidence, the notion that *P. vivax* is benign is unsustainable (Cowman et al. 2016).

The most common symptom of malaria infection in persons without natural or acquired immunity is fever. After treatment, fever will usually subside over a few days. In severe cases, malaria can lead to anemia, cerebral malaria (coma), and respiratory distress (Cowman et al. 2016). However, in a population with stable malarial infection, immunity increases with age, with the proportion of severe cases negligible after age 10, and asymptomatic infection being the dominant infection type beyond age 15 (Cowman et al. 2016).

### Lifecycle

Part of the ongoing interest in modelling malaria is its complicated lifecycle. Malaria is a vector borne disease, needing both human (or other vertebrate) and mosquito hosts to complete its lifecycle. Malaria first enters the human blood stream via the skin after the female mosquito has a blood meal. From the blood stream it proceeds to the liver where it proliferates and is released into the blood where it enters the red blood cells and reproduces further. Eventually, the parasites undergo sexual differentiation, maturing in the bone marrow until they are released into the blood stream to be consumed by a mosquito during a blood feed where it matures into sporozoites ready to reinfect a new vertebrate host (Cowman et al. 2016).

Figure 3.1 depicts this lifecycle for *P. vivax* malaria, which has an additional stage in its lifecycle to *P. falciparum*. *P. vivax* malaria can form hypnozoites, a dormant liver stage of the parasite. These can remain dormant for weeks and even months, leading to recurrent infections and illness, possibly until the conditions for transmission are more favourable. In subtropical/temperate areas, the incubation periods can be between 8-12 months, compared to 3-4 weeks in tropical regions. Price et al. 2020. *P. vivax* also has lower levels of the blood stage parasite during infection, which

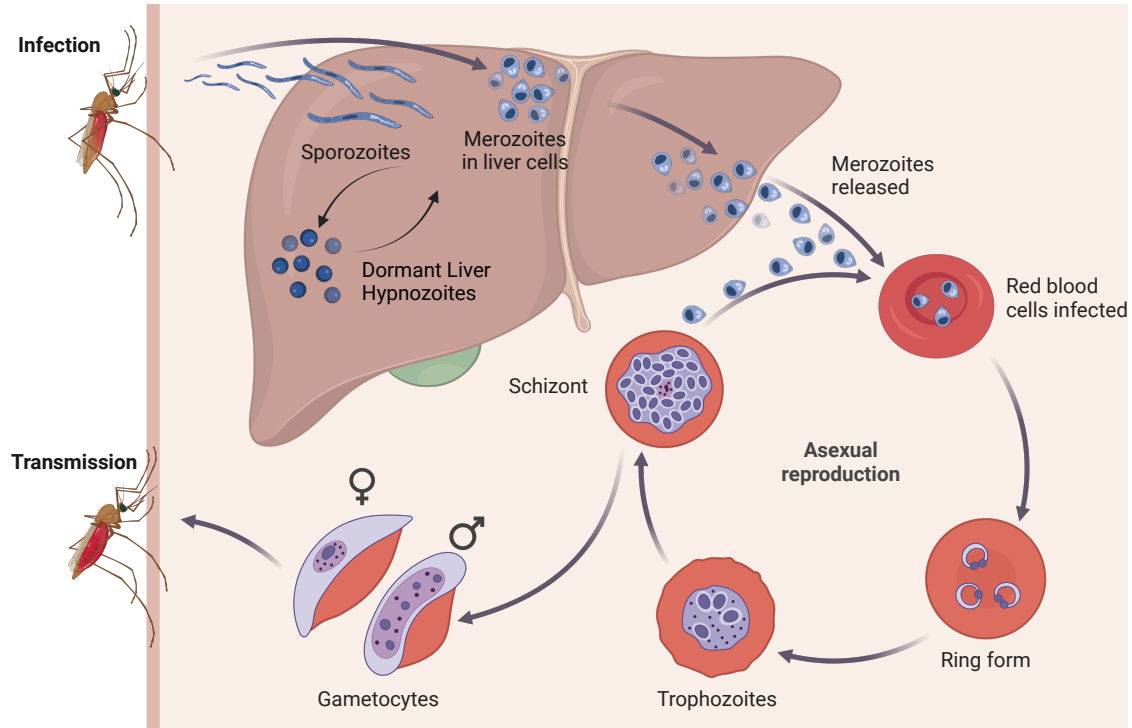


Figure 3.1: The *P. vivax* (malaria) lifecycle. *P. falciparum* does not have a dormant liver hypnozoite stage. Created with BioRender.com.

means diagnosis is more difficult, and it has an increased proportion of asymptomatic cases (Adams and Mueller 2017).

### 3.1 Mathematical Modelling of Malaria

Levels of asymptomatic cases and dormant parasite (in the case of *P. vivax*) are impossible or difficult to experimentally determine without mass testing. By creating a model of the disease, and calibrating the model so that it simulates symptomatic case levels reported by health authorities, it is possible to estimate these previously ‘hidden’ levels. Furthermore, malaria models allow us to simulate the effects of public health interventions such as mass treatment or testing. Malaria models also allow us to estimate how economical interventions are likely to be, before large amounts of money are spent on trials.

#### Ross-Macdonald

The most basic model capturing the fundamental lifecycle of malaria is commonly referred to as the Ross-Macdonald model. One such example of such a model was used in Aron and May 1982, and is depicted in Figure 3.2. It accounts for human to mosquito and mosquito to human transmission by having compartments for susceptible and infected humans ( $S_H$  and  $I_H$ ), as well as susceptible

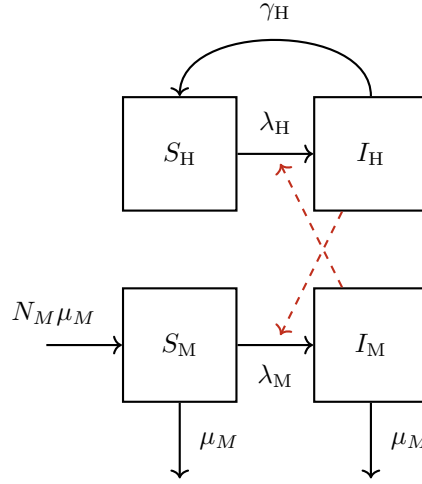


Figure 3.2: A simple Ross-Macdonald malaria model schematic, as described by Aron and May 1982.  $S_H$  and  $I_H$  are the number of susceptible and infected humans respectively, and  $S_M$  and  $I_M$  are the number of susceptible and infected mosquitos. The rate of human infection ( $\lambda_H$ ) is dependant on  $I_M$ , and the rate of human infection ( $\lambda_M$ ) is dependant on  $I_H$ .

and infected mosquitos ( $S_M$  and  $I_M$ ). The ODEs for this model are

$$\begin{aligned}\frac{dS_H}{dt} &= \gamma_H I_H - b T_{HM} I_M \frac{S_H}{N_H} \\ \frac{dI_H}{dt} &= b T_{HM} I_M \frac{S_H}{N_H} - \gamma_H I_H \\ \frac{dS_M}{dt} &= N_M \mu_M + \gamma_M I_M - b T_{MH} S_M \frac{I_H}{N_H} - S_M \mu_M \\ \frac{dI_M}{dt} &= b T_{MH} S_M \frac{I_H}{N_H} - \gamma_M I_M\end{aligned}$$

where  $b$  is the biting rate per mosquito, and  $T_{HM}$  is the probability of transmission to a human given a bite by an infectious mosquito, with  $T_{MH}$  being vice-versa. Note that it is  $\frac{I_H}{N_H}$  in the mosquito dynamics. Biologically this is assuming the number of blood meals a mosquito takes per day is invariant to the size of the human population. Mosquitos do not have a possibility of leaving the infected stage apart from death due to their short lifespans, but the births and deaths are mathematically equivalent to assuming that the rate of ‘recovery’ amongst mosquitos is  $\mu_M I_M$  per unit time, with no population dynamics.

The broader class of Ross-Macdonald style models simplify the lifecycle of malaria to the following four steps (Smith et al. 2012):

1. Malaria is transmitted to human (or vertebrate) via a blood feed.
2. Malaria proliferates in the human host until it circulates in the peripheral blood
3. A mosquito then takes a blood feed, ingesting the pathogen
4. Malaria develops within the mosquito host, progressing to its salivary glands, able to infect a human.

## Models of *P. Vivax* Malaria

The basic Ross-Macdonald model does not sufficiently capture the dynamics of *P. vivax*, since it does not consider relapses. Therefore *P. vivax* specific models have been proposed which introduce compartments for dormant liver stage infection.

### White Model

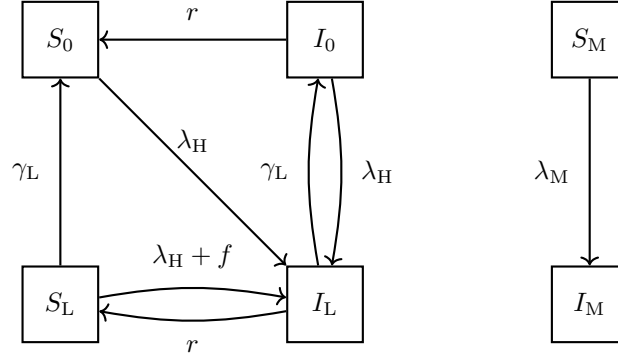


Figure 3.3: Diagram for *P. vivax* model in a tropical setting described by White et al. 2016.  $S$  and  $I$  are the number of susceptible and infected humans and mosquitos (denoted by subscript M).  $\lambda_H = mabI_M$  and  $\lambda_M = ac(I_0 + I_L)$

One such model that incorporates a dormant stage is the White model for *P. vivax* depicted in Figure 3.3 (White et al. 2016, Tropical Model). The model is comprised of six compartments:

1.  **$S_0$  (Susceptible Individuals - No Latent Hypnozoite Liver Stage Infection)** : People in this compartment have no form of malarial infection. These people are susceptible to new malarial infections, and are infected into compartment  $I_L$  (with both blood and liver stage parasites) with rate  $\lambda_H$ .
2.  **$I_L$  (Infected Individuals - Both Blood Stage and Latent Hypnozoite Liver Stage Infection)** : Individuals in this compartment have both an active blood-stage infection, and latent hypnozoite infection in the liver. Individuals in  $I_L$  can progress to either  $I_0$  through the clearance of liver stage infection with rate  $\gamma_L$ , or to  $S_L$  through clearance of blood stage infection with rate  $r$ .
3.  **$I_0$  (Infected Individuals - Blood-Stage Infection Only)** : Individuals in  $I_0$  have a blood-stage infection with no latent hypnozoite infection in the liver. They are be reinfected into  $I_L$  with rate  $\lambda_H$ , relapse with rate  $f$ . Blood-stage infection is cleared (moving into compartment  $S_0$ ) with rate  $r$ .
4.  **$S_L$  (Susceptible Individuals - Blood-Stage Infection Only)** : Individuals in  $S_L$  have latent hypnozoite infection in the liver without blood-stage infection. They get novel infection through a mosquito bite into  $I_L$  with rate  $\lambda_H$ , or hypnozoite activation with rate  $f$ . This means that those in  $S_L$  move to compartment  $I_L$  with total rate  $\lambda_H + f$ . Alternatively the hypnozoites are cleared from the liver (moving to compartment  $S_0$ ) with rate  $\gamma_L$ .
5.  **$S_M$  (Susceptible Mosquitoes)** : Susceptable mosquitoes become infectious at rate  $\lambda_M p$ . They die at rate  $g + \lambda_M(1 - p)$ . Since there is a constant mosquito population assumption, mosquitoes are born into this state at rate  $g + \lambda_M$ .



6.  $I_M$  (**Infectious Mosquitoes**): Infectious mosquitos die at rate  $g + \lambda_M(1 - p)$ .

The White model is characterised by the following ODEs:

$$\begin{aligned}\frac{dS_0}{dt} &= -\lambda_H S_0 + rI_0 + \gamma_L S_L \\ \frac{dI_0}{dt} &= -\lambda_H I_0 - rI_0 + \gamma_L I_L \\ \frac{dS_L}{dt} &= -\lambda_H S_L + rI_L - fS_L - \gamma_L S_L \\ \frac{dI_L}{dt} &= \lambda_H(S_0 + I_0 + S_L) - rI_L + fS_L - \gamma_L I_L \\ \frac{dS_M}{dt} &= g - \lambda_M(pS_M - (1 - p)I_M) - gS_M \\ \frac{dI_M}{dt} &= \lambda_M(pS_M - (1 - p)I_M) - gI_M. \quad (I_0 + I_L = \text{total number of bloodstage infections})\end{aligned}$$

Note that here each compartment is expressed as a proportion of the population.

The force of infection for humans is defined as  $\lambda_H := mabI_M$  where  $m$  is the number of mosquitos per human (held constant since there is no birth or death in the human dynamics),  $a$  is the mosquito biting rate, and  $b$  is the probability that a human bitten by an infectious mosquito develops an infection.

The force of infection for mosquitos is defined as  $\lambda_M := ac(I_0 + I_L)$  where  $a$  is defined above, and  $c$  is the probability that a mosquito bite on an infectious mosquito causes the mosquito to become infectious.  $g$  can be interpreted as the natural birth/death rate for mosquitos.  $p$  is then the proportion of mosquitos that survive long enough after the initial infection that the parasite matures enough in the mosquito before becoming infectious to new susceptible humans. Under the assumption that time until parasite transmissability after infection in a mosquito is a constant  $n$  days, and that mosquitoes naturally die at rate  $g$ ,  $p = e^{-gn}$ . To see this let  $V \sim \text{Exp}(g)$ , represent the lifespan of the mosquito.  $\Pr(V > n) = 1 - F_V(n) = 1 - (1 - e^{-gn}) = e^{-gn}$ .

$\lambda_M(1 - p)$  can be interpreted as an additional rate of death, where of the mosquitos that would develop malaria after a bite, a proportion  $1 - p$  die instantly. This applies to both the susceptible and infectious mosquitos. Presumably this approximates a model where mosquitoes are moved to an ‘exposed’ compartment for  $n$  time, after initial infection, however no justification is given by White et al. for this additional parameter  $n$ . A more straightforward  $SI$  model could be constructed that absorbs  $c$  and  $n$  into the single parameter  $c^*$ , such that it becomes the proportion of mosquito bites on blood stage infectious humans that result in mosquito infection where the mosquito does not die before becoming infectious. With steady mosquito population, the mosquito dynamics would now be characterised by

$$\frac{dI_M}{dt} = \lambda_M^* S_M - gI_M \quad \text{where } \lambda_M^* := ac^*(I_0 + I_L).$$

By modelling both liver and bloodstage infection, blood stage infections from relapses can be captured in the dynamics, meaning it is possible to analyse case number data that may be confounded by relapses as well as novel infections.

This model does not account for continual depletion of liver stage parasites which would vary the rate of relapse over time (through clearance or relapse). It also does not directly model any

interventions or case importations. The lack of population dynamics means the model may only be useful on a small time scale. Finally, it does not account for any importation of disease from an outside area, so if  $S_0 = 1$ , *P. vivax* is presumed permanently eradicated.

### Champagne Model

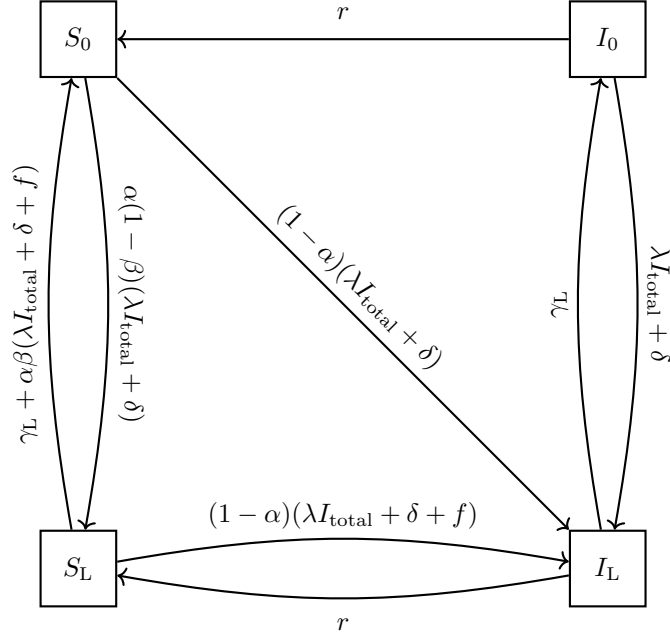


Figure 3.4: Diagram for *P. vivax* model described by Champagne et al. 2022.  $I_{\text{total}} = I_0 + I_L$ . Since the mosquito dynamics have been removed,  $\lambda$  now not has no dependencies on the number of infectious mosquitos.

The Champagne model - described in (Champagne et al. 2022) and diagrammatically depicted in figure 3.4 - both simplifies and extends the White model. The model assumes human to human transmission, removing mosquito dynamics, and extends it by adding in a rate of imported cases and treatment of malarial infection. It is characterised by the system of ODEs:

$$\begin{aligned} \frac{dI_L}{dt} &= (1 - \alpha)(\lambda I_{\text{total}} + \delta)(S_0 + S_L) + (\lambda I_{\text{total}} + \delta)I_0 + (1 - \alpha)fS_L - \gamma_L I_L - rI_L \\ \frac{dI_0}{dt} &= -(\lambda I_{\text{total}} + \delta)I_0 + \gamma_L I_L - rI_0 \\ \frac{dS_L}{dt} &= -(1 - \alpha(1 - \beta))(\lambda I_{\text{total}} + \delta + f)S_L + \alpha(1 - \beta)(\lambda I_{\text{total}} + \delta)S_0 - \gamma_L S_L + rI_L \\ \frac{dS_0}{dt} &= -(1 - \alpha\beta)(\lambda I_{\text{total}} + \delta)S_0 + (\lambda I_{\text{total}} + \delta)\alpha\beta S_L + \alpha\beta fS_L + \gamma_L S_L + rI_0 \end{aligned}$$

where  $I_{\text{total}} := I_0 + I_L$ .

The compartments  $S_0, I_0, I_L$  and  $S_L$  have the same interpretation as in the White model, however the rates between compartments are significantly modified.

The new parameters are

- $\lambda$  : the rate of infection
- $\delta$  : importation rate

- $\alpha$  : proportion of those infected who clear blood stage infections through immediate treatment, and
- $\beta$  : proportion of those cleared of blood stage infection who are also cleared of liver stage parasites (radical cure).

In other words, the proportion of infected individuals  $\alpha\beta$  are completely cured from liver and blood stage parasites. The model assumes treatment clears infection instantaneously. Individuals in  $S_L$  who relapse or get a new infection are assumed to be cured with the same proportions as new infections from  $S_0$ , but individuals in  $I_0$  who are superinfected are assumed not to seek treatment.

In contrast to the White model, the Champagne model allows analysis of potential treatment interventions, or how much of an impact limiting the importation rate might have on case numbers (through border control/testing). Although the lack of mosquito dynamics simplifies the model and it's running, it is unrealistic. The model still has some of the same problems as the White model, such as not incorporating hypnozoite depletion rates and a lack of population dynamics, meaning all analytic results are done assuming the system is at equilibrium.



# Chapter 4

## Parameter Inference

### 4.1 Motivation

Building mathematical models of real world phenomenon allows for us to simulate changes in the world without having to undertake large scale experiments. However, once we have a model that sufficiently approximates *P. vivax* transmission or anything else we are trying to model, we then need to estimate what the ‘true’ underlying parameters are. To do this we calibrate the model against real world data such as case counts, and prevalence surveys. Under frequentist assumptions, there is a ‘true’ set of parameters that if used in our model, simulated the observed data. Under a Bayesian assumption, the parameters are considered to be random, and This chapter explores statistical inference techniques to recover the parameters, under both the frequentist and Bayesian frameworks.

### 4.2 Frequentist Parameter Estimation

Assume the model is parametrised by a set of parameters  $\theta \in \Theta$  which we are trying to estimate by considering some observed data  $\mathbf{y}^{\text{obs}} = (y_1^{\text{obs}}, \dots, y_n^{\text{obs}})$ . Consider  $\mathbf{y}(\theta) = (y_1(\theta), \dots, y_n(\theta))$  some model simulation of  $\mathbf{y}^{\text{obs}}$ . Often the observed data has some underlying index set  $x_1, \dots, x_n$ , where  $x_i$  might be something like time. In this case we can also consider the observed data to be  $\{(x_1, y_1^{\text{obs}}), \dots, (x_n, y_n^{\text{obs}})\}$ , and the model simulated data to be  $\{(x_1, y_1(\theta)), \dots, (x_n, y_n(\theta))\}$ .

#### Least Squares Estimator

It is common that models are not random, but instead model the mean behaviour of a system. In this case,  $\mathbf{y}(\theta)$  is not random. Therefore we can assume that  $y_i^{\text{obs}} = y_i(\theta) + \varepsilon_i$ , where  $\varepsilon_i$  is a random variable with some (possibly unknown) distribution, and zero mean.

When the distribution of  $\varepsilon_i$  is unknown, a common approach for estimating  $\theta^{\text{LSE}}$  is to take the least squares estimate.

**Definition 4.1** (Least Squares Estimate). *The least squares estimate  $\theta^{\text{LSE}}$  for  $\theta$  is*

$$\theta^{\text{LSE}} := \arg \min_{\theta \in \Theta} \sum_{i=1}^n (y_i(\theta) - y_i^{\text{obs}})^2.$$

**Example 4.2.** Consider the observed data  $\{(x_1, y_1^{\text{obs}}), (x_2, y_2^{\text{obs}}), (x_3, y_3^{\text{obs}})\} = \{(1, 2), (2, 4), (3, 4)\}$ , which we assume were generated from the model  $y_i(\boldsymbol{\theta}) + \varepsilon_i$ , where  $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ , and  $\mathbb{E}(\varepsilon_i) = 0$ . We derive the least squares estimate of our parameters  $\boldsymbol{\theta} = (\theta_0, \theta_1)$  by

$$\begin{aligned}\boldsymbol{\theta}^{\text{LSE}} &= \arg \min_{\boldsymbol{\theta}} \left[ \sum_{i=1}^3 (y_i(\boldsymbol{\theta}) - y_i^{\text{obs}})^2 \right] \\ &= \arg \min_{\boldsymbol{\theta}} \left[ \sum_{i=1}^3 (\theta_0 + \theta_1 x_i - y_i^{\text{obs}})^2 \right] \\ &= \arg \min_{\boldsymbol{\theta}} [(\theta_0 + \theta_1 - 2)^2 + (\theta_0 + 2\theta_1 - 4)^2 + (\theta_0 + 3\theta_1 - 4)^2].\end{aligned}$$

Since the expanded quadratic will have positive coefficients for  $\theta_0$  and  $\theta_1$ , we can solve for  $\boldsymbol{\theta}^{\text{LSE}}$  by

$$\begin{aligned}\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\theta}} [(\theta_0^{\text{LSE}} + \theta_1^{\text{LSE}} - 2)^2 + (\theta_0^{\text{LSE}} + 2\theta_1^{\text{LSE}} - 4)^2 + (\theta_0^{\text{LSE}} + 3\theta_1^{\text{LSE}} - 4)^2] \\ &= \begin{bmatrix} 6\theta_0^{\text{LSE}} + 12\theta_1^{\text{LSE}} - 20 \\ 12\theta_0^{\text{LSE}} + 28\theta_1^{\text{LSE}} - 44 \end{bmatrix}\end{aligned}$$

And solving these two equations results in  $\theta_0^{\text{LSE}} = 4/3$  and  $\theta_1^{\text{LSE}} = 1$ . This can be visually seen in Figure 4.1 in the red line that minimises the sum of the squares of the difference between the observations (in black dots) and the linear model.

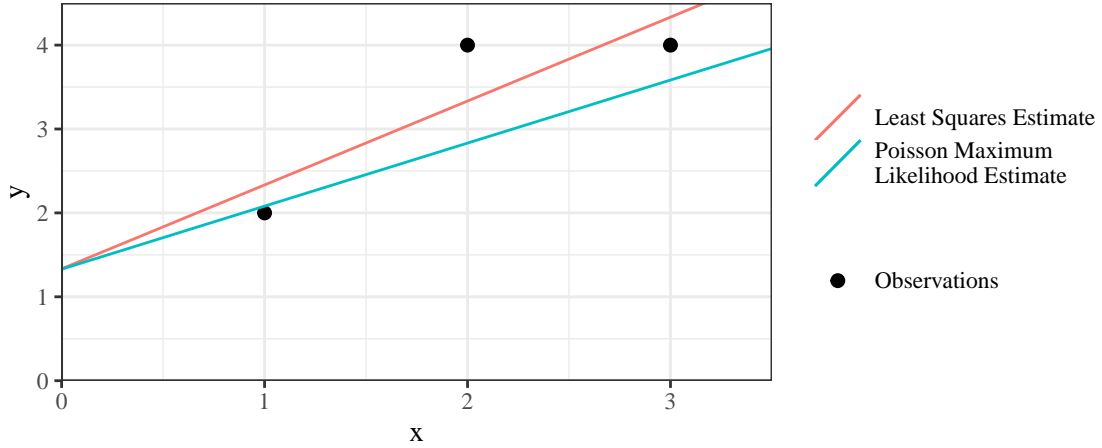


Figure 4.1: Two linear models of the form  $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$  fit given the set of observations  $\{(1, 2), (2, 4), (3, 4)\}$  using the method of least squares and maximum likelihood under the assumption that the data are independent realisations from a Poisson distribution with mean  $y_i(\boldsymbol{\theta})$ . The least squares estimates were  $\theta_0^{\text{LSE}} = 4/3$  and  $\theta_1^{\text{LSE}} = 1$ . The maximum likelihood estimates were  $\hat{\theta}_0 \approx 1.329$  and  $\hat{\theta}_1 \approx 0.751$ .

## Maximum Likelihood Estimator

The least squares method makes no explicit assumptions about the distribution of the noise  $\varepsilon$ . However if the distribution of  $\varepsilon$  is known (or can be reasonably assumed), we can explicitly calculate

the probability of the data given the parameters.

**Definition 4.3** (Likelihood function). *With  $\mathbf{y}^{\text{obs}}$  fixed, the likelihood function is*

$$\mathcal{L}(\boldsymbol{\theta}) := \Pr(\mathbf{y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} = \mathbf{y}^{\text{obs}} | \boldsymbol{\theta}).$$

*Particularly, if  $y_i(\boldsymbol{\theta}) + \varepsilon_i$  are independent*

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n \Pr(y_i(\boldsymbol{\theta}) + \varepsilon_i = y_i^{\text{obs}} | \boldsymbol{\theta}).$$

The dependence of the likelihood function on  $\mathbf{y}^{\text{obs}}$  is notationally suppressed, but can be explicitly written as  $\mathcal{L}(\boldsymbol{\theta} | \mathbf{y}^{\text{obs}})$  to avoid confusion. In the continuous (or mixture of discrete and continuous) case, we interpret  $\Pr(\mathbf{y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} = \mathbf{y}^{\text{obs}} | \boldsymbol{\theta})$  as the density  $\Pr(\mathbf{y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \in d\mathbf{y}^{\text{obs}} | \boldsymbol{\theta})$  with respect to an underlying probability measure.

A natural estimate for  $\boldsymbol{\theta}$  is the one that maximises the likelihood function  $\mathcal{L}$ , as it coincides with the value of  $\boldsymbol{\theta}$  that maximises the probability of the data (possibly expressed as a density). Such an estimate is called the maximum likelihood estimate.

**Definition 4.4** (Maximum Likelihood Estimate). *The maximum likelihood estimate of  $\boldsymbol{\theta}$  is*

$$\hat{\boldsymbol{\theta}} := \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}).$$

It is often computationally easier to deal with the log-likelihood  $\ell(\boldsymbol{\theta}) := \ln \mathcal{L}(\boldsymbol{\theta})$ . Since the natural logarithm is a monotonic function,

$$\arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}).$$

**Example 4.5.** *Using the same observed data set as Example 4.2, we assume that  $y_i^{\text{obs}}$  were generated independently from  $y_i(\boldsymbol{\theta}) + \varepsilon_i \sim \text{Pois}(y_i(\boldsymbol{\theta}))$ , where  $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$  as previously defined. Therefore the maximum likelihood estimate of  $\boldsymbol{\theta}$  is*

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^3 y_i^{\text{obs}} \ln(y_i(\boldsymbol{\theta})) - y_i^{\text{obs}}(\boldsymbol{\theta}) - \ln(y_i^{\text{obs}}!) \\ &= \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^3 y_i^{\text{obs}} \ln(\theta_0 + \theta_1 x_i) - \theta_0 - \theta_1 x_i - \ln(y_i^{\text{obs}}!) \\ &= \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} 2 \ln(\theta_0 + \theta_1) - \theta_0 - \theta_1 + 4 \ln(\theta_0 + 2\theta_1) - \theta_0 - 2\theta_1 + 4 \ln(\theta_0 + 3\theta_1) - \theta_0 - 3\theta_1 \end{aligned}$$

*which we can solve to get  $\hat{\theta}_0 \approx 1.329$  and  $\hat{\theta}_1 \approx 0.751$ , as seen in Figure 4.1 with the green line. This estimate is not the same as the linear model estimate using the least squares estimates for  $\boldsymbol{\theta}$*

## Relationship of Least Squares and Maximum Likelihood Estimates

Although the least squares estimate does not explicitly assume a distribution, it coincides with the maximum likelihood estimate under the assumption that the  $y_i^{\text{obs}}$  were generated with i.i.d. normal error.

**Theorem 4.6.** *If  $y_i(\boldsymbol{\theta}) + \epsilon_i \sim N(y_i(\boldsymbol{\theta}), \sigma^2)$ , then*

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{\text{LSE}}.$$

*Proof.*

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(y_i(\boldsymbol{\theta}) - y_i^{\text{obs}})^2}{\sigma^2} \\ &= \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n -\frac{(y_i(\boldsymbol{\theta}) - y_i^{\text{obs}})^2}{\sigma^2} \\ &= \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n -(y_i(\boldsymbol{\theta}) - y_i^{\text{obs}})^2 \\ &= \arg \min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n (y_i(\boldsymbol{\theta}) - y_i^{\text{obs}})^2 \\ &= \boldsymbol{\theta}^{\text{LSE}}. \end{aligned}$$

□

Surprisingly, the proof only requires  $\sigma^2$  to be a constant, but does not require it to be known.

## Frequentist Parameter Estimates in Compartmental Models

Various approaches are possible to parameterise compartmental models. If the stochastic compartmental model is simple enough, and the number of people in the model is small enough, then the likelihood for the stochastic model could be calculated directly. However this is hardly ever the case, and approximations are usually made.

For a model with a single unknown parameter, such as in it is possible to fit a deterministic ODE model fit to a single data point. For example, Champagne et al. 2022 fits one unknown model parameter to incidence data. Fitting to a single data point is not generally advisable, as the parameter estimates will not be robust to variations in the observation of that data. Alternatively, if there are multiple observations to fit the model to parameters can be estimated by finding the least squares estimates fit to the ODE model. Gani and Leach 2001 fit part of their modified *SEIR* smallpox model using least squares estimates. Another approach is to assume that the observed data follow a particular distribution determined by the ODE solution. For example, it is plausible to assume that daily incidence (case counts) in an *SIS* model such as described by Equations 2.1 and 2.2, could be distributed according to a Poisson distribution, with a mean number of cases  $\beta \frac{I_t}{N} S_t$ , where  $I_t$  and  $S_t$  are solutions of the ODEs at time  $t$ . This is because  $\beta \frac{I_t}{N} S_t$  describes the rate at which individuals are moving from the susceptible compartment to the infected compartment. Other data such as samples from the population to estimate the proportion of population who are infectious could be distributed according to  $\text{Binom}(n, \frac{I_t}{N})$ , where  $n$  is the total number of people sampled.

Figure 4.2 demonstrates estimation of one unknown parameter  $\beta$  from an *SEIR* model, using prevalence data. Both  $\beta^{\text{LSE}}$  in grey and  $\hat{\beta}$  in yellow are very close, but do not fit the data well suggesting fitting the ODEs to a stochastic model simulation may be a poor choice. This is because



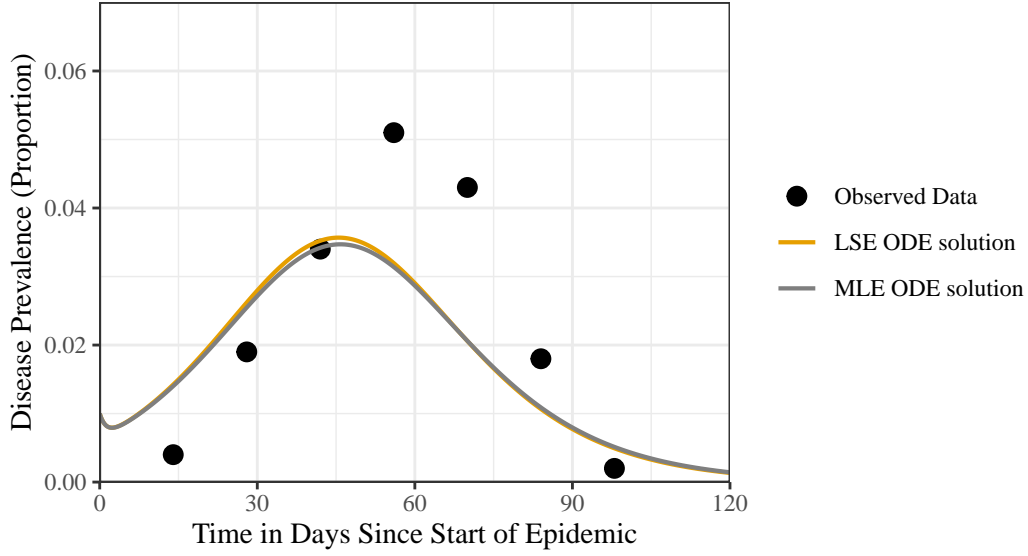


Figure 4.2: An *SEIR* model fit to some observed prevalence data taken every two weeks over a 14 week period, generated as  $I_t/N$  from the *SEIR* simulation in Figure 2.3. All parameters were considered known except for  $\beta$ . The least squares estimate (LSE)  $\beta^{\text{LSE}} = 0.3516$  was found by solving the model ODEs and numerically minimising the square differences between observed prevalences and the ODE prevalences (as proportions). Similarly the maximum likelihood estimate  $\hat{\beta} = 0.3493$  was found by assuming the prevalence (times 1000) was binomially distributed from 1000 samples with the probability of being infectious equal to  $\frac{I_t}{N}$ .

at the beginning of an epidemic the behaviour is very stochastic. Therefore trying to fit an ODE model to it's stochastic analogue is not necessarily a good idea. The ODEs are more likely to well approximate the stochastic model when the number of people in each compartment is high.

### 4.3 Bayesian Parameter Estimation

In frequentist statistical inference,  $\theta$  is considered to be fixed, with the observed data  $\mathbf{y}^{\text{obs}}$  assumed to be generated from a distribution depending on  $\theta$ . Although it is possible to quantify the uncertainty in parameter estimates through confidence intervals, frequentist estimates naturally lend themselves to point estimates. In contrast, inference under a Bayesian framework assumes that  $\theta$  is also a random variable that before any observations are made follows a prior distribution. ‘Evidence’ from the observed data then updates belief about  $\theta$ , resulting in a posterior distribution of  $\theta$ , described by Bayes’ theorem, namely

$$\Pr(\theta|\mathbf{y}^{\text{obs}}) \propto \Pr(\mathbf{y}^{\text{obs}}|\theta) \Pr(\theta).$$

Bayesian parameter estimation is still dependent on the likelihood function  $\mathcal{L}(\theta) := \Pr(\mathbf{y}^{\text{obs}}|\theta)$ . By using the posterior estimates of our parameter values, we can generate a posterior predictive distribution of our model. This can be helpful in forecasting, as it can capture uncertainty more robustly (at least than point estimates). For instance, a government may be interested in the number of additional hospital beds that need to be available to cope with an outbreak of a disease. If a disease model can be used to approximate outbreaks of the disease, we can use previous instances of the disease to calibrate our model parameters. Samples from the posterior parameter

distribution, allow the model to be run multiple times with varying sets of parameters, and provide a range of predicted outcomes for the disease. This allows for confidence in how much investment may be required in the health system. Similarly, samples from the posterior parameter distribution allow for scenario modelling such as introducing a new vaccine.

If  $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}})$  is a known distribution with established sampling methods, we can sample directly from the distribution. If not we have to use other methods to sample from the distribution.

## Rejection Sampling

---

### Algorithm 3 Rejection Sampler

---

```

Sample  $\Theta^* \sim p$ 
Sample  $U \sim \text{Unif}(0, 1)$ 
if  $U \leq \frac{g(\Theta^*)}{Mp(\Theta^*)}$  then
    return  $\Theta^*$  as a sample from the distribution of  $\Theta$ 
else
    Reject  $\Theta^*$  as a sample from the distribution of  $\Theta$ , and repeat the algorithm
end if

```

---

Rejection sampling is not just a method for sampling from a posterior distribution, but rather a method for sampling from any distribution where the density can be calculated up to a proportionality constant. Algorithm 3 demonstrates how this can be done for  $\Theta$  with density proportional to  $g$ . Given a distribution  $p$  and constant  $M$  such that  $Mp(x) \geq g(x)$ , we sample a proposal  $\Theta^*$  from  $p$ . We then accept  $\Theta^*$  as a sample from  $g$  with probability  $\frac{g(\Theta^*)}{Mp(\Theta^*)}$ , and rejected otherwise.

Under this methodology, let  $\Theta^*$  be the proposed sampled. The distribution function of the accepted samples  $\Theta$  is

$$\begin{aligned}
 \Pr(\Theta = \theta) &\propto \Pr\left(\Theta^* = x, U \leq \frac{g(\Theta^*)}{Mp(\Theta^*)}\right) \\
 &\quad \text{(where the probabilities may be interpreted as densities)} \\
 &= \Pr\left(U \leq \frac{g(X^*)}{Mp(X^*)} | X^* = x\right) \Pr(X^* = x) \\
 &= \frac{g(x)}{Mp(x)} p(x) \\
 &= \frac{g(x)}{M}
 \end{aligned}$$

as required.

**Example 4.7.** Let  $g(x) = (x - 1)^2$  be an unnormalised density function for  $x \in (0, 2)$ . It can be shown that  $g(x) \leq 1$ , the density of a  $\text{Unif}(0, 1)$  random variable. Therefore to generate samples from  $g$  we sample uniformly from  $X^* \sim \text{Unif}(0, 1)$ , and accept the sample if a new  $U \sim \text{Unif}(0, 1)$  is less than  $(X^* - 1)^2$ . This is demonstrated in Figure 4.3, with the black line being the unnormalised density, the green dots being accepted samples, and red dots being rejected samples. The values along the  $X^*$  axis correspond to samples from  $g$ .

## Markov Chain Monte Carlo Methods

Often it is not possible to sample directly from the posterior distribution  $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}})$  using a rejection sampler, as there is no explicit form proportional to the true density. Therefore a common

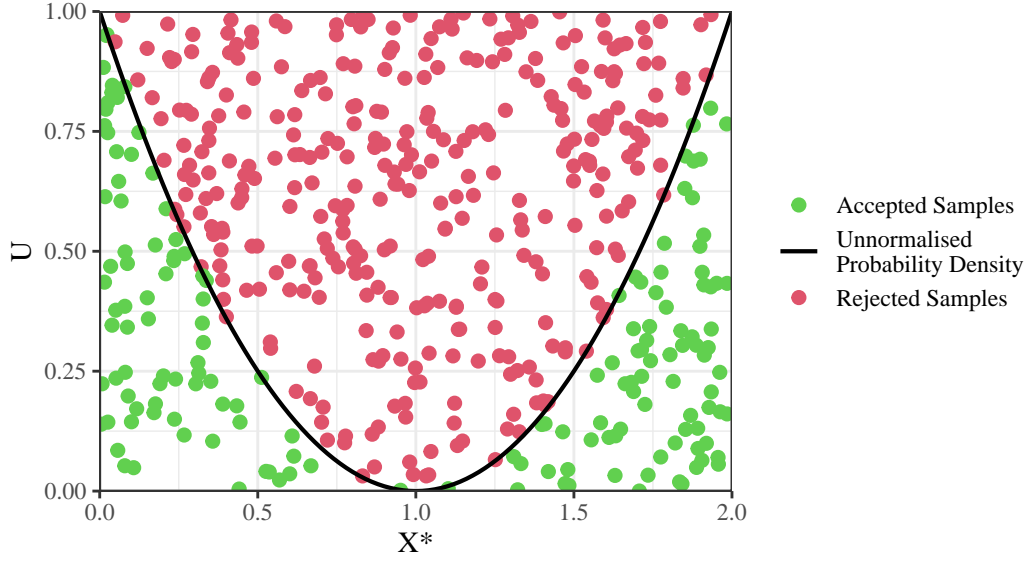


Figure 4.3: Samples of  $X$  from the unnormalised density  $g(x) = (x - 1)^2$  with  $x \in (0, 2)$  using the rejection sampler.  $X^* \sim \text{Unif}(0, 2)$  and  $M = 1$ . Green dots are samples from  $X$ . Of 500 samples of  $X^*$ , 157 were accepted as samples of  $X$ .

way of sampling from a distribution  $p(x)$  is to construct a Markov chain with stationary distribution  $p(x)$ . Hence, eventually each new state the chain moves to will be a (not necessarily independent) sample from  $p(x)$ , or in our case  $\Pr(\theta | \mathbf{y}^{\text{obs}})$ .

**Definition 4.8** ((Discrete-Time) Markov Chain). *A sequence of random variables  $X_0, X_1, \dots$  is a (discrete-time) Markov chain  $\{X_i\}_{i \in \mathbb{N}}$  if for all  $k \in \mathbb{N}$ ,*

$$\Pr(X_{i+1} \in A | X_0, X_1, \dots, X_i) = \Pr(X_{i+1} \in A | X_i).$$

If  $X_i \in \mathcal{X}$ , then  $\mathcal{X}$  is the state space of the Markov chain. For example the Markov chain depicted in Figure 4.4, has discrete state space  $\mathcal{X} = \{1, 2\}$ , and transition probabilities depicted by the numbers on the arrows.

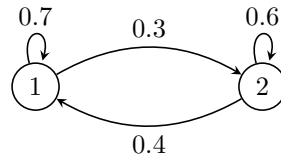


Figure 4.4: A simple time homogeneous Markov chain, with two states. It is characterised by the transition kernel  $K(1, 1) = \Pr(X_{i+1} = 1 | X_i = 1) = 0.7$ ,  $K(1, 2) = \Pr(X_{i+1} = 2 | X_i = 1) = 0.3$ ,  $K(2, 1) = \Pr(X_{i+1} = 1 | X_i = 2) = 0.4$ , and  $K(2, 2) = \Pr(X_{i+1} = 2 | X_i = 2) = 0.6$ . The stationary distribution is  $\pi(1) = 4/7$  and  $\pi(2) = 3/7$ .

Markov chains are characterised by a transition kernel  $K$  with

$$K(x_i, x_{i+1}) := \Pr(X_{i+1} = x_{i+1} | X_i = x_{i+1}),$$

where this probability is interpreted as a density for continuous random variables.  $K(1, 1)$  would therefore be the probability of transitioning from state 1 to state 2 in a single discrete time step.

We will restrict our focus to time homogeneous Markov chains where if the value of  $X_i$  is known to be  $x$ , behaviour of chain from this point on will be identical to the behaviour of the chain from  $X_j$ , if this is also observed to be  $x$ .

**Definition 4.9** (Time Homogeneous). *A Markov chain is time homogeneous if*

$$\{X_i, X_{i+1}, \dots, X_{i+n}\} \stackrel{d}{=} \{X_{i'}, X_{i'+1}, \dots, X_{i'+n}\}$$

for all  $i, i', n \in \mathbb{N}$ , given  $X_i = x = X_{i'}$ .

The Markov chain in Figure 4.4 is time homogeneous. It does not matter how long it took to get into a state, the Markov chain will behave the same from that point forward.

**Definition 4.10** (Stationary Distribution). *A Markov chain has stationary distribution  $\pi$  if for  $X_i \sim \pi$ , then  $X_{i+1}|X_i \sim \pi$ .*

**Example 4.11.** *Given the Markov chain in Figure 4.4, the stationary distribution can be calculated by solving the simultaneous equations*

$$\begin{aligned} K(1, 1) \times \pi(1) + K(2, 1) \times \pi(2) &= 0.7 \times \pi(1) + 0.4 \times \pi(2) = \pi(1) \\ \pi(1) + \pi(2) &= 1. \end{aligned}$$

Therefore  $\pi(1) = 4/7$  and  $\pi(2) = 3/7$ .

As stated earlier, to sample from a distribution  $p(x)$ , we construct a Markov chain with this stationary distribution. A sufficient condition to know that we have achieved this is if our chain satisfies the detailed balance condition.

**Theorem 4.12** (Detailed balance condition). *A Markov chain has stationary distribution  $p(x)$ , which it converges to independent of initialisation, if for all  $x, x'$ ,*

$$p(x)K(x, x') = p(x')K(x', x).$$

*Proof.* More formally this requires the notions of recurrent, nonnull, irreducible and aperiodic Markov chains which we do not discuss here. For a full discussion and proof see Robert and Casella 2010, Chapter 6.  $\square$

### Metropolis-Hastings

The Metropolis-Hastings algorithm is one way of constructing a Markov chain with stationary distribution equal to the target distribution  $g$ . We choose a proposal distribution  $q(x'|x)$  which given our last sample  $x$ , generates a new random variable  $X'$ . For example  $q$  might be the density of  $X' \sim N(x, 1)$ , a normal random variable with mean around the previous sample. Then similar to rejection sampling,  $X'$  is accepted as the next state in the distribution with some probability  $\alpha$ , chosen in such a way that if the chain is distributed according to a stationary distribution  $X_i \sim g$ , then the next step will also be distributed according to that stationary distribution  $X_{i+1} \sim g$ . Formally this is set out in Algorithm 4.

Note that for symmetric proposal distributions  $q(x'|x) = q(x|x')$ ,  $\alpha$  simplifies to  $\min\left(\frac{g(x')}{g(x)}, 1\right)$ , in which case the algorithm is simply called a Metropolis sampler.

**Algorithm 4** Metropolis-Hastings Sampler

---

```

Initialise  $x_0$ 
for  $i = 1$  to  $N$  do
  Sample  $X' \sim q(x'|x_{i-1})$ 
  Compute acceptance ratio  $\alpha = \min\left(\frac{g(x')q(x_{i-1}|x')}{g(x_{i-1})q(x'|x_{i-1})}, 1\right)$ 
  Sample  $U \sim \text{Uniform}(0, 1)$ 
  if  $U \leq \alpha$  then
     $x_i \leftarrow X'$ 
  else
     $x_i \leftarrow x_{i-1}$ 
  end if
end for
return  $\{x_0, x_1, \dots, x_N\}$ 

```

---

**Theorem 4.13.** *The chain produced by Algorithm 4  $\{X_k\}_{k \in \mathbb{N}}$  has stationary distribution  $g$  for proposal distributions that cover the support of  $g$ .*

*Proof.* We show that the detailed balance condition

$$g(x)q(x'|x)\alpha(x, x') = g(x')q(x|x')\alpha(x', x)$$

where  $\alpha(x, x') = \min\left(\frac{g(x')q(x|x')}{g(x)q(x'|x)}, 1\right)$  is satisfied. Without loss of generality let  $g(x')q(x|x') < g(x)q(x'|x)$ , so  $\alpha(x, x') = \frac{g(x')q(x|x')}{g(x)q(x'|x)}$ , and  $\alpha(x', x) = 1$ .

$$\begin{aligned}
 g(x)q(x'|x)\alpha(x, x') &= g(x)q(x'|x) \times \frac{g(x')q(x|x')}{g(x)q(x'|x)} \\
 &= g(x')q(x|x') \\
 &= g(x')q(x|x')\alpha(x', x)
 \end{aligned}$$

□

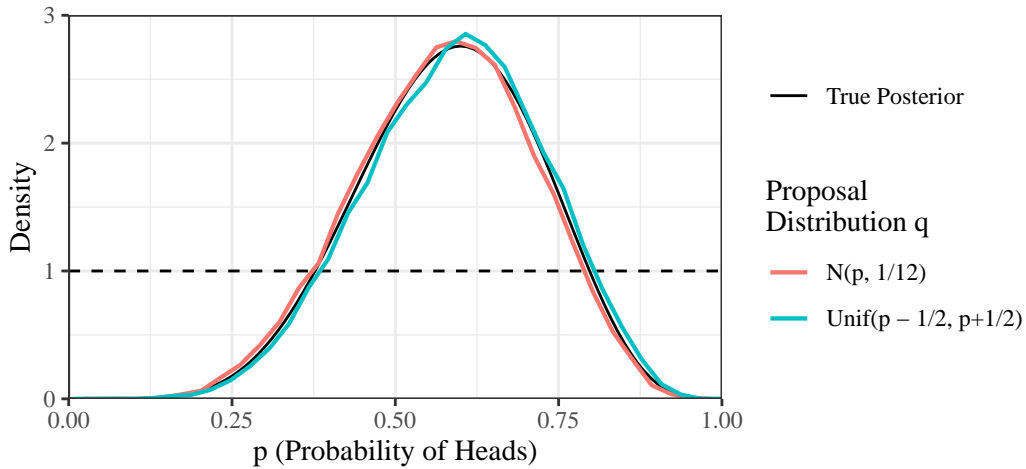


Figure 4.5: Samples from the posterior distribution of  $p$  using the Metropolis-Hastings algorithm.  $p$  was assumed to have a uniform prior between 0 and 1, with  $y^{\text{obs}} = 6$ , generated from  $\text{Binom}(10, p)$ . The choice of proposal distribution did not impact the final estimate of  $\Pr(p|y^{\text{obs}})$ .

As seen empirically in Example 4.14 and Figure 4.5, the proof does not depend on choice of proposal distribution.

**Example 4.14** (Coin toss). *Let the probability of tossing a heads on a weighted coin be  $\Pr(X = 1) = p$ . Assume that  $p \sim \text{Unif}(0, 1)$ . We observe  $y^{\text{obs}} = 6$  heads from 10 tosses of the coin. Therefore*

$$\Pr(p|y^{\text{obs}}) \propto \Pr(y^{\text{obs}}|p) \Pr(p) = \binom{n}{y^{\text{obs}}} p^{y^{\text{obs}}} (1-p)^{n-y^{\text{obs}}} \times 1 = 210p^6(1-p)^4.$$

We sample from this distribution using the Metropolis algorithm which becomes

```

Initialise  $p_0$ 
for  $i = 1$  to  $N$  do
  Sample  $P' \sim q(p'|p_{i-1})$ 
  Compute acceptance ratio  $\alpha = \min\left(\frac{(P')^6(1-P')^4}{p_{i-1}^6(1-p_{i-1})^4}, 1\right)$   $\triangleright$  Assuming  $q$  symmetric
  Sample  $U \sim \text{Uniform}(0, 1)$ 
  if  $U \leq \alpha$  then
     $p_i \leftarrow P'$ 
  else
     $p_i \leftarrow p_{i-1}$ 
  end if
end for
return  $\{p_0, p_1, \dots, p_N\}$ 

```

We can compare two different proposal distributions for  $q(p'|p)$ ,  $P' \sim N(p, 1/12)$ , and the second being  $P' \sim \text{Unif}(p - 1/2, p + 1/2)$ . The first 1000 samples were discarded as burn in, and it was thinned to every 5 samples. The resulting distribution of the samples can be seen in Figure 4.5, with both proposal distributions resulting in samples that are good at estimating the true distribution.

Since the chain converges to the stationary distribution over time, and is highly correlated, a derived chain  $\{x_{B+iT}\}_{i \in \mathbb{N}}$  is constructed from the output. The first  $B$  samples are discarded as ‘burn in’ samples to reduce the impact of the initialisation point. The chain is ‘thinned’ by taking every  $T$ th sample, since  $X_i$  and  $X_{i+1}$  may also be highly correlated (in samples where the proposed  $X'$  is rejected,  $X_i = X_{i+1}$ ). This derived chain is considered a random sample from the target distribution  $g$ . In practice, diagnostics such as trace plots and autocorrelation plots are used to determine  $B$  and  $T$  (see Gelman et al. 2014, Chapter 11).

For disease models, given a prior distribution for the parameter(s)  $\theta$ , Metropolis-Hastings can be used to produce samples from  $\Pr(\theta|y^{\text{obs}}) \propto \mathcal{L}(\theta) \Pr(\theta)$ , where  $\mathcal{L}(\theta) \Pr(\theta)$  can be calculated to a proportionality constant but not directly sampled from. For example given an *SIS* model described in Equations 2.1 and 2.2 with unknown effective contact rate  $\beta \sim \text{Gamma}(2, 6)$  and daily case counts  $y^{\text{obs}}$ , we can draw samples from  $\theta|y^{\text{obs}}$  as in Figure 4.6. This figure demonstrates that the choice of likelihood results in different posterior sample distributions. The samples from Poisson likelihood have greater variance than the samples from the binomial likelihood.

## Gibbs Sampling

Some models, may have a parameters such that it is possible to sample from  $\theta_1|\theta_2, \mathbf{y}^{\text{obs}}$ , and  $\theta_2|\theta_1, \mathbf{y}^{\text{obs}}$  but not the joint distribution of  $(\theta_1, \theta_2)|\mathbf{y}^{\text{obs}}$ . A (multidimensional) Markov chain can be constructed by iteratively updating the parameters. Such a method is called a Gibbs sampler,

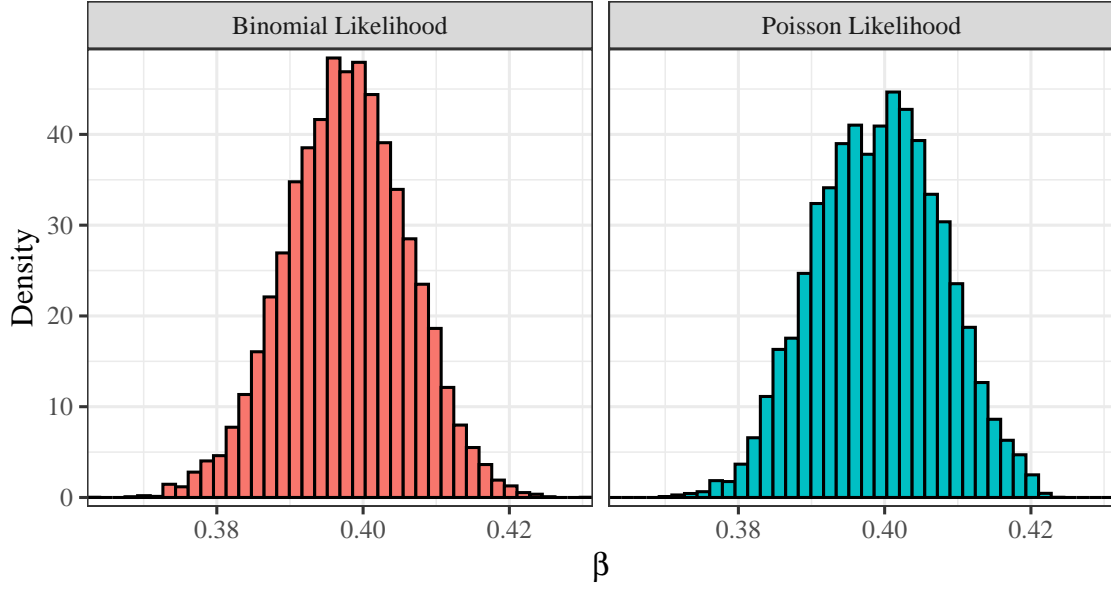


Figure 4.6: Given a daily incidence of  $y^{\text{obs}} = 26$  at day 30 of an *SIS* epidemic, with unknown  $\beta$ , we use Metropolis-Hasting to sample from  $\Pr(\beta|y^{\text{obs}})$ .  $\gamma = 1/4$  was assumed to be correct, and we compared the assumption  $y^{\text{obs}} \sim \text{Binom}(\lfloor S_{30} \rfloor, \beta I_{30}/N)$ , to the assumption  $y^{\text{obs}} \sim \text{Pois}(\frac{\beta I_{30} S_{30}}{N})$  where  $I_{30}, S_{30}$  are the ODE solutions to Equations 2.1 and 2.2. We assumed the prior distribution  $\beta \sim \text{Gamma}(2, 6)$ , where  $\mathbb{E}(\beta) = 1/3$ . Our proposal density was  $N(\beta^*, 1/10)$ , where  $\beta^*$  was the previous sample.

---

**Algorithm 5** Gibbs Sampler
 

---

```

Initialise  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})$ 
for  $i = 1$  to  $N$  do
  Sample  $\theta_1^{(i)} \sim \Pr(\theta_1 | \theta_2^{(i-1)}, \theta_3^{(i-1)}, \dots, \theta_d^{(i-1)})$ 
  Sample  $\theta_2^{(i)} \sim \Pr(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \dots, \theta_d^{(i-1)})$ 
   $\vdots$ 
  Sample  $\theta_d^{(i)} \sim \Pr(\theta_d | \theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_{d-1}^{(i)})$ 
  Save  $(\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_d^{(i)})$  as  $\theta^{(i)}$ 
end for
return  $\{\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(N)}\}$ 
  
```

---

described in Algorithm 5. The distribution of the Markov chain sampler will eventually converge to the  $\Pr(\theta_1, \theta_2 | \mathbf{y}^{\text{obs}})$ , and for the same reasons as for the Metropolis-Hastings sampler, after thinning and discarding burn in, we consider the resulting chain a sequence of independent samples from our target distribution.

**Theorem 4.15** (Gibbs Sampler). *The Markov chain generated by Algorithm 5 converges to the distribution of  $\Pr(\boldsymbol{\theta} | \mathbf{y}^{\text{obs}})$ .*

*Proof.* We prove that the Gibbs Sampler satisfies the detailed balance equation for two unknown parameters. The transition kernel of the Markov chain is

$$q(\boldsymbol{\theta}^{(i)} | \boldsymbol{\theta}^{(i-1)}) := \Pr(\theta_1^{(i)} | \theta_2^{(i-1)}, \mathbf{y}) \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}).$$

To prove the detailed balance condition is satisfied, we need to show that

$$\Pr(\boldsymbol{\theta}^{(i-1)} | \mathbf{y}) q(\boldsymbol{\theta}^{(i)} | \boldsymbol{\theta}^{(i-1)}) = \Pr(\boldsymbol{\theta}^{(i)} | \mathbf{y}) q(\boldsymbol{\theta}^{(i-1)} | \boldsymbol{\theta}^{(i)}).$$

$$\begin{aligned} \Pr(\boldsymbol{\theta}^{(i-1)} | \mathbf{y}) q(\boldsymbol{\theta}^{(i)} | \boldsymbol{\theta}^{(i-1)}) &= \Pr(\theta_1^{(i-1)}, \theta_2^{(i-1)} | \mathbf{y}) \times \Pr(\theta_1^{(i)} | \theta_2^{(i-1)}, \mathbf{y}) \times \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}) \\ &= \Pr(\theta_1^{(i-1)} | \theta_2^{(i-1)}, \mathbf{y}) \times \Pr(\theta_2^{(i-1)} | \mathbf{y}) \times \Pr(\theta_1^{(i)} | \theta_2^{(i-1)}, \mathbf{y}) \times \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}) \\ &= \Pr(\theta_1^{(i-1)} | \theta_2^{(i-1)}, \mathbf{y}) \times \Pr(\theta_1^{(i)}, \theta_2^{(i-1)} | \mathbf{y}) \times \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}) \\ &= \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}) \times \Pr(\theta_1^{(i)}, \theta_2^{(i-1)} | \mathbf{y}) \times \Pr(\theta_1^{(i-1)} | \theta_2^{(i-1)}, \mathbf{y}) \\ &= \Pr(\theta_2^{(i)} | \theta_1^{(i)}, \mathbf{y}) \times \Pr(\theta_1^{(i)} | \mathbf{y}) \times \Pr(\theta_2^{(i-1)} | \theta_1^{(i)}, \mathbf{y}) \times \Pr(\theta_1^{(i-1)} | \theta_2^{(i-1)}, \mathbf{y}) \\ &= \Pr(\theta_1^{(i)}, \theta_2^{(i)} | \mathbf{y}) \times \Pr(\theta_2^{(i-1)} | \theta_1^{(i)}, \mathbf{y}) \times \Pr(\theta_1^{(i-1)} | \theta_2^{(i-1)}, \mathbf{y}) \\ &= \Pr(\boldsymbol{\theta}^{(i)} | \mathbf{y}) q(\boldsymbol{\theta}^{(i-1)} | \boldsymbol{\theta}^{(i)}) \end{aligned}$$

as required, so the posterior distribution  $\Pr(\boldsymbol{\theta} | \mathbf{y})$  is the unique stationary distribution associated with the generated Markov chain.  $\square$

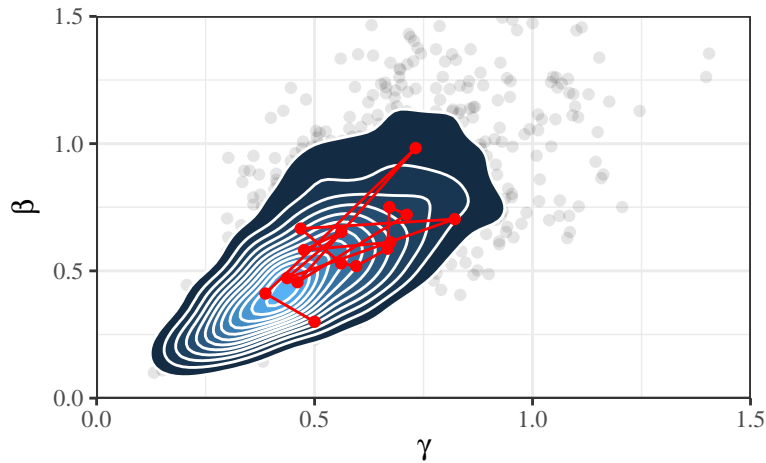


Figure 4.7: 2000 posterior samples from  $\Pr(\beta, \gamma | \mathbf{y}^{\text{obs}})$ , where  $\beta | \gamma, \mathbf{y}^{\text{obs}} \sim \text{Gamma}(9, 4/\gamma + 4 + 8\gamma)$  and  $\gamma | \beta, \mathbf{y}^{\text{obs}} \sim \text{InvGamma}(12, 12\beta)$ . The samples were obtained using a Gibbs sampler. The red points are the first 15 samples using the Gibbs sampler.



**Example 4.16.** Consider the SIS model described by Equations 2.1 and 2.2. Early in an epidemic, the average number of new cases generated from a single infectious individual is known as  $R_0$ . This can be shown to be  $\frac{\beta}{\gamma}$  for the SIS model. Let  $\mathbf{y}^{\text{obs}} = \{1, 1, 3, 1\}$  be the number of people infected by four different individuals at the start of the epidemic. We assume that the number of infections are generated from a Poisson distribution with mean  $\frac{\beta}{\gamma}$ . Therefore the likelihood is

$$\mathcal{L}(\beta, \gamma) := \Pr(\mathbf{y}^{\text{obs}} | \beta, \gamma) = \frac{\left(\frac{\beta}{\gamma}\right)^{1+1+3+1} \exp(-\frac{4\beta}{\gamma})}{1! \times 1! \times 3! \times 1!} \propto \left(\frac{\beta}{\gamma}\right)^6 \exp(-\frac{4\beta}{\gamma}).$$

Let us assume that from similar previous epidemics we assume that  $\beta | \gamma \sim \text{Gamma}(3, 4 + 8\gamma)$ , and  $\gamma | \beta \sim \text{InvGamma}(6, 8\beta)$ . Therefore

$$\begin{aligned} \Pr(\beta | \gamma, \mathbf{y}^{\text{obs}}) &\propto \Pr(\mathbf{y}^{\text{obs}} | \gamma, \beta) \Pr(\beta | \gamma) \\ &\propto \left(\frac{\beta}{\gamma}\right)^6 \exp(-\frac{4\beta}{\gamma}) \times \beta^{3-1} \exp(-(4 + 8\gamma)\beta) \\ &\propto \beta^{9-1} \exp(-(4/\gamma + 4 + 8\gamma)\beta) \end{aligned}$$

and so  $\beta | \gamma, \mathbf{y}^{\text{obs}} \sim \text{Gamma}(9, 4/\gamma + 4 + 8\gamma)$ . Similarly

$$\begin{aligned} \Pr(\gamma | \beta, \mathbf{y}^{\text{obs}}) &\propto \Pr(\mathbf{y}^{\text{obs}} | \gamma, \beta) \Pr(\gamma | \beta) \\ &\propto \left(\frac{\beta}{\gamma}\right)^6 \exp(-\frac{4\beta}{\gamma}) \times \gamma^{-6-1} \exp\left(-\frac{8\beta}{\gamma}\right) \\ &\propto \gamma^{-12-1} \exp\left(-\frac{12\beta}{\gamma}\right) \end{aligned}$$

and so  $\gamma | \beta, \mathbf{y}^{\text{obs}} \sim \text{InvGamma}(12, 12\beta)$ . Now we have explicit forms for the conditional probabilities, we generate samples using the Gibbs sampler in Algorithm 5. Samples from the distribution can be seen in Figure 4.7.

The Gibbs sampler and Metropolis-Hastings sampler are often combined, by using a Metropolis-Hastings sampler for each step of the conditional sampling. This is useful when the conditional distributions  $\Pr(\theta_1 | \theta_2, \mathbf{y}^{\text{obs}})$  can be calculated up to a proportionality constant, but not directly sampled from.

## Approximate Bayesian Computation

So far, under a Bayesian framework, parameter estimation has still been dependent on the likelihood function  $\mathcal{L}(\theta) := \Pr(\mathbf{y}^{\text{obs}} | \theta)$  through Bayes' theorem. In many cases, such as stochastic disease models and agent based models, the likelihood has no explicit form, or is intractable to calculate. The only option here is to run the model given  $\theta$ , and sample  $\mathbf{y}(\theta)$  directly.

---

### Algorithm 6 Naive Bayesian Sampler

---

```

Sample  $\theta^* \sim \Pr(\theta)$ 
Run model and compute  $\mathbf{y}(\theta^*)$ 
if  $\mathbf{y}(\theta^*) = \mathbf{y}^{\text{obs}}$  then
    return  $\theta^*$  as a sample from  $\Pr(\theta | \mathbf{y}^{\text{obs}})$ 
end if

```

---

A naive method of using such model runs to sample from  $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}})$  is to sample  $\boldsymbol{\theta}^*$  from  $\Pr(\boldsymbol{\theta})$ , and run the model to obtain  $\mathbf{y}(\boldsymbol{\theta}^*)$ . For each iteration,  $\mathbf{y}(\boldsymbol{\theta}^*)$  will exactly equal  $\mathbf{y}^{\text{obs}}$  with probability  $\Pr(\mathbf{y}^{\text{obs}}|\boldsymbol{\theta}^*)\Pr(\boldsymbol{\theta}^*) \propto \Pr(\boldsymbol{\theta}^*|\mathbf{y}^{\text{obs}})$ , and hence if  $\mathbf{y}(\boldsymbol{\theta}^*) = \mathbf{y}^{\text{obs}}$  we can accept  $\boldsymbol{\theta}^*$  as a sample from our posterior parameter distribution. This is outlined in Algorithm 6. When  $\mathbf{y}^{\text{obs}}|\boldsymbol{\theta}^*$  does not have a countable number of non-zero probability outputs,  $\mathbf{y}(\boldsymbol{\theta}^*) = \mathbf{y}^{\text{obs}}$  can be exactly zero, and even in the countable case, for higher dimensional  $\mathbf{y}(\boldsymbol{\theta})$ , the probability of returning exactly  $\mathbf{y}^{\text{obs}}$  vanishes. Therefore we draw inspiration from the continuous interpretation of the likelihood  $\mathcal{L}(\boldsymbol{\theta}) := \Pr(\mathbf{y}(\boldsymbol{\theta}) \in d\mathbf{y}^{\text{obs}}|\boldsymbol{\theta})$ . Since

$$\Pr(\mathbf{y}(\boldsymbol{\theta}) \in d\mathbf{y}^{\text{obs}}|\boldsymbol{\theta}) := \lim_{\epsilon \rightarrow 0} \frac{\Pr(\mathbf{y}(\boldsymbol{\theta}) \in B_\epsilon^D(\mathbf{y}^{\text{obs}}))}{\epsilon}, \quad (4.1)$$

where  $B_\epsilon^D(\mathbf{y}^{\text{obs}})$  is a ball of size  $\epsilon$  around  $\mathbf{y}^{\text{obs}}$ , with respect to some (unknown) metric  $D$  induced by the (unknown) probability distribution of  $\mathbf{y}(\boldsymbol{\theta})$ . Therefore  $\mathcal{L}(\boldsymbol{\theta})$  is approximately proportional to  $\Pr(\mathbf{y}(\boldsymbol{\theta}) \in B_\epsilon^D(\mathbf{y}^{\text{obs}})|\boldsymbol{\theta})$ , (as a function of  $\boldsymbol{\theta}$ ).

Hence we construct a new approximate sampling algorithm where rather than rejecting the sample for  $\mathbf{y}(\boldsymbol{\theta}) \neq \mathbf{y}^{\text{obs}}$ , we accept  $\mathbf{y}(\boldsymbol{\theta})$  if it falls within a ball of size  $\epsilon$  around  $\mathbf{y}^{\text{obs}}$ . Equivalently we accept samples if  $D(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}}) < \epsilon$ .

Since the distribution of  $\mathbf{y}(\boldsymbol{\theta})$  is unknown, and since the metric required for Equation 4.1 to hold depends on  $\boldsymbol{\theta}$ , we do not explicitly derive  $D(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}})$ , we are forced to approximate it with  $\tilde{D}(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}})$ . The most common choice of  $\tilde{D}$  is the  $L^p$  norm of  $\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}^{\text{obs}}$

$$\tilde{D}(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}}) = \|\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}^{\text{obs}}\|_p = \left( \sum_{i=1}^d |y_i(\boldsymbol{\theta}) - y_i^{\text{obs}}|^p \right)^{1/p},$$

for  $p \geq 1$ . For  $p = 1$  or  $2$  this is the Manhattan or Euclidean distance between the two vectors. When the observations in  $\mathbf{y}^{\text{obs}}$  are on different scales, or  $\mathbf{y}^{\text{obs}}$  are highly correlated between model runs, care needs to be taken to rescale  $\mathbf{y}^{\text{obs}}$  and  $\mathbf{y}(\boldsymbol{\theta})$  by a covariance matrix to remove correlation, or by rescaling using the relative differences instead of  $\|\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}^{\text{obs}}\|_p$ .

---

**Algorithm 7** Approximate Bayesian Computation Sampler

---

```

Sample  $\boldsymbol{\theta}^* \sim \Pr(\boldsymbol{\theta})$ 
Run model and compute  $\mathcal{D}(\boldsymbol{\theta}^*)$ 
if  $\mathcal{D}(\boldsymbol{\theta}^*) < \epsilon$  then
    return  $\boldsymbol{\theta}^*$  as a sample from  $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}})$ 
end if

```

---

The full procedure is outlined in Algorithm 7, but since  $\mathbf{y}^{\text{obs}}$  is fixed, we consider  $\tilde{D}(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}})$  as a function of  $\boldsymbol{\theta}$ , and we can equivalently write  $\mathcal{D}(\boldsymbol{\theta}) := \tilde{D}(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}})$ . We call  $\mathcal{D}(\boldsymbol{\theta})$  the discrepancy function where in a non-deterministic model,  $\mathcal{D}(\boldsymbol{\theta})$  is random.

## Chapter 5

# Gaussian Processes and Synthetic Likelihoods

If the distribution of  $\mathcal{D}(\boldsymbol{\theta})$  was known for all  $\boldsymbol{\theta}$ , then the need to sample from the model within Algorithm 7 would be redundant, since  $\mathcal{D}(\boldsymbol{\theta})$  could be sampled from directly. Moreover, the probability of drawing and accepting a sample using this algorithm becomes  $\Pr(\boldsymbol{\theta}) \Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta})$ . Comparing this to Bayes' theorem, we can see that  $\Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta})$  plays the role of the likelihood. Therefore  $L(\boldsymbol{\theta}) := \Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta})$  is an approximation of  $\mathcal{L}(\boldsymbol{\theta})$  (up to some proportionality constant). In reality, we do not know the distribution of  $\mathcal{D}(\boldsymbol{\theta})$ , for all  $\boldsymbol{\theta}$ , but we consider methods to construct an approximation  $\hat{L}$  of  $L$  which we call the synthetic likelihood. The approximation considered is achieved by modelling  $\mathcal{D}(\boldsymbol{\theta})$  using a surrogate model which we introduce this chapter.

### 5.1 Gaussian Processes

For  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$  that are close to each other, the distribution of  $\mathcal{D}(\boldsymbol{\theta})$  will be similar to the distribution of  $\mathcal{D}(\boldsymbol{\theta}')$ . Therefore sampling from  $\mathcal{D}(\boldsymbol{\theta})$  gives information about the distribution of  $\mathcal{D}(\boldsymbol{\theta}')$  for  $\boldsymbol{\theta}, \boldsymbol{\theta}'$  close. A reasonable assumption could be that  $\mathbb{E}(\mathcal{D}(\boldsymbol{\theta}_1)), \mathbb{E}(\mathcal{D}(\boldsymbol{\theta}_2)), \dots, \mathbb{E}(\mathcal{D}(\boldsymbol{\theta}_n))$  are multivariate normally distributed with  $\text{cov}(E(\mathcal{D}(\boldsymbol{\theta}_1)), E(\mathcal{D}(\boldsymbol{\theta}_2)))$  large for  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$  close, and small for  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$  far apart, so that  $\text{cov}(E(\mathcal{D}(\boldsymbol{\theta}_1)), E(\mathcal{D}(\boldsymbol{\theta}_2)))$  is a function of  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$ . Formally, we treat  $\mathbb{E}(\mathcal{D}(\beta))$  as a realisation of a Gaussian process.

**Definition 5.1** (Gaussian Process). *A collection of random variables  $\{f(x)\}_{x \in \mathcal{X}}$  (where  $x$  may be a vector) is a Gaussian process if any finite subset of the collection of random variables is multivariate normal distributed. That is, there is a function  $m : \mathcal{X} \rightarrow \mathbb{R}$  and symmetric kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  such that for all finite sets  $\mathbf{x} := \{x_1, x_2, \dots, x_n\} \subset \mathcal{J}$ , with  $f(\mathbf{x}) := [f(x_1), f(x_2), \dots, f(x_n)]^T$*

$$f(\mathbf{x}) \sim \text{MVN} \left( \begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_n, x_1) & \dots & \dots & k(x_n, x_n) \end{bmatrix} \right).$$

**Definition 5.2** (Mean Function and Covariance Kernel). *The mean function and covariance kernel*

are

$$m(x_i) := \mathbb{E}[f(x_i)]$$

and

$$k(x_i, x_{i'}) := \text{cov}(f(x_i), f(x_{i'})).$$

Gaussian processes are simultaneously realised over the whole space  $\mathcal{X}$  (for example  $\mathbb{R}^d$ ) and are hence collections of (uncountably infinite) random variables. However the covariance function  $k$  is always constructed such that  $\text{corr}(x, x') \rightarrow 1$  as  $\|x - x'\| \rightarrow 0$ . This induces a form of continuity in  $x$  called mean square continuity which is defined in Definition 5.4 The most famous example of a Gaussian process is Brownian motion.

**Definition 5.3** (Brownian Motion).  $B(t) : \mathbb{R}^+ \rightarrow \mathbb{R}$  is a Brownian motion on  $\mathbb{R}$  if

1.  $B(0) = 0$  almost surely
2.  $B(t_0), B(t_1) - B(t_0), \dots, B(t_n) - B(t_{n-1})$  are independent for all  $t_0 < t_1 < t_2 < \dots < t_n$
3.  $B(t + s) - B(t) \sim N(0, s)$  for  $s, t \geq 0$
4.  $B(t)$  is continuous almost surely for  $t > 0$ .

Brownian motion has zero mean, covariance kernel  $k(s, t) = \min(s, t)$ , which implies that  $\text{corr}(B_s, B_t) = \frac{\min(s, t)}{\sqrt{st}}$ . This is close to 1 as  $s$  is close to  $t$ , however as  $|s - t| \rightarrow \infty$ ,  $\text{corr}(B_s, B_t) \rightarrow 0$ .

Gaussian processes can be considered as giving a probability distribution to functions  $f \in \mathcal{F}$ , where  $\mathcal{F}$  is a class of functions. The properties of the class of functions  $\mathcal{F}$  depend on the covariance kernel  $k$ . Different  $k$  result in very different functions. One such property to consider is the smoothness of the functions, which we describe by mean square differentiability.

**Definition 5.4** (Mean Square Continuous). A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is mean square continuous at  $\mathbf{x}$  in the  $i$ th direction at  $\mathbf{x}$  if  $\mathbb{E}(|f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})|^2) \rightarrow 0$  as  $|h| \rightarrow 0$ , where  $\mathbf{e}_i$  is the unit vector with a 1 in the  $i$ th coordinate.

**Definition 5.5** (Mean Square Differentiable). A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is mean square differentiable at  $\mathbf{x}$  in the  $i$ th direction with derivative  $\frac{\partial f(\mathbf{x})}{\partial x_i}$  if

$$\mathbb{E} \left[ \left| \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \frac{\partial f(\mathbf{x})}{\partial x_i} \right|^2 \right] \rightarrow 0$$

as  $|h| \rightarrow 0$ , where  $\mathbf{e}_i$  is the unit vector in the direction of the  $i$ th coordinate.

The concept of mean square differentiability and continuity are analogous to differentiability and continuity in the non-random function case.

**Theorem 5.6.** Brownian motion is mean square continuous, but not mean square differentiable.

*Proof.*  $(B_{t+h} - B_t)^2 \sim (\sqrt{|h|}Z)^2$  where  $Z \sim N(0, 1)$ . Therefore  $(B_{t+h} - B_t)^2 \sim |h|\chi_1^2 \rightarrow 0$  almost surely as  $|h| \rightarrow 0$ , hence  $\mathbb{E}[(B_{t+h} - B_t)^2] = 0$ , and so Brownian motion is mean square continuous. Since  $\frac{B_{t+h} - B_t}{h} \sim N(0, 1/|h|)$ ,  $\frac{B_{t+h} - B_t}{h}$  does not converge to any valid probability distribution as  $|h| \rightarrow 0$ , as the variance approaches  $+\infty$ , and so Brownian motion is not mean square differentiable.  $\square$

## Kernels

### Matérn Kernel Family

Since we are motivated to consider  $\mathbb{E}(\mathcal{D}(\boldsymbol{\theta}))$  as a realisation of a Gaussian process, we consider some common choices of kernel function and their effect on the Gaussian process realisations. The two most common families of kernel functions are the squared exponential and Matérn families (Rasmussen and Williams 2008). The Matérn kernel explicitly allows for adjustment of function smoothness through a hyperparameter  $\nu$ .

**Definition 5.7** (Matérn Kernel).

$$k_\nu(x, x') = \sigma_k^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu} \|x - x'\|}{\ell} \right)^\nu K_\nu \left( -\frac{\sqrt{2\nu} \|x - x'\|}{\ell} \right)$$

where  $K_\nu$  is a modified Bessel function (defined in Abramowitz and Stegun 2013, p. 374).  $\nu, \ell$ , and  $\sigma_k$  are hyperparameters.

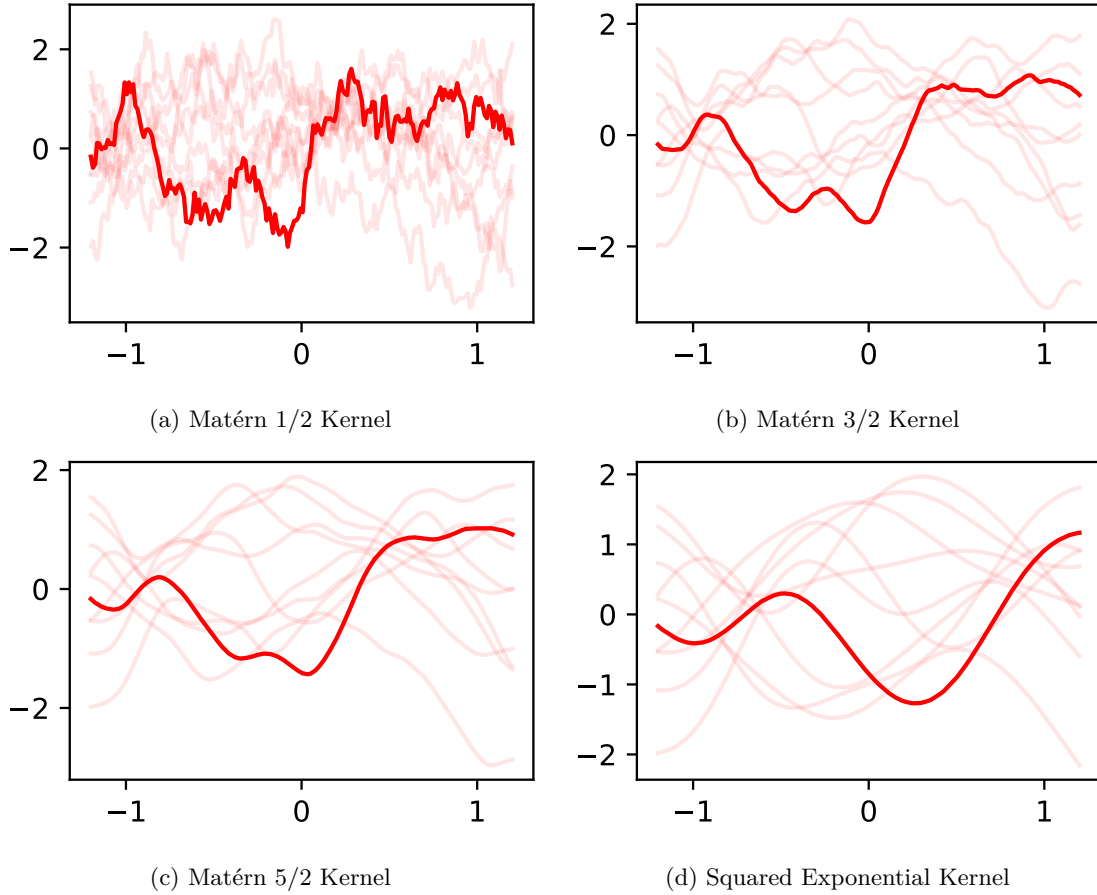


Figure 5.1: Ten sample realisations from 4 different kernels, with one realisation bolded. Samples for each kernel were generated from the same seed and the hyperparameters  $\ell$ , and  $\sigma_k$  were set to 1.

Realisations from zero mean Gaussian processes with this kernel are  $\lfloor \nu \rfloor$  times mean square differentiable (Rasmussen and Williams 2008). The most common values for  $\nu$  are  $1/2, 3/2$  and

5/2, which result in functions that are 0, 1, and 2 times mean square differentiable. In these cases the kernel can be slightly simplified to:

$$k_{1/2}(x, x') = \sigma_k^2 \exp\left(-\frac{\|x - x'\|}{\ell}\right),$$

$$k_{3/2}(x, x') = \sigma_k^2 \left(1 + \frac{\sqrt{3}\|x - x'\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|}{\ell}\right),$$

and

$$k_{5/2}(x, x') = \sigma_k^2 \left(1 + \frac{\sqrt{5}\|x - x'\|}{\ell} + \frac{5\|x - x'\|^2}{3\ell^2}\right) \exp\left(-\frac{\|x - x'\|^2}{2 * \ell^2}\right).$$

The increasing smoothness of these kernels can be seen in Figures 5.1a, 5.1b, and 5.1c.

Zero mean Gaussian processes with a Matérn kernel are  $n$  times mean square differentiable, for all  $n < \nu$ . As seen in Figure 5.1, this means that this kernel allows for flexibility in how smooth realised functions are.

### Squared Exponential Kernel

As  $\nu \rightarrow \infty$ , the Matérn kernel converges to a kernel which we call the squared exponential kernel (Rasmussen and Williams 2008, p. 85).

**Definition 5.8** (Squared Exponential Kernel).

$$k(x, x') = \sigma_k^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right)$$

By construction, the squared exponential kernel is infinitely mean square differentiable, which can visually be seen in 5.1d. The squared exponential kernel is considered the default kernel in much of the literature,

Despite this being the ‘default’ kernel in much of the literature (for example Gutmann and Cor 2016), infinite differentiability is a very strong condition that may not be appropriate in all circumstances.

### Length and Amplitude Hyperparameters

Both the Matérn and squared quadratic kernels (as well as most other common kernels choices), there are two hyperparameters  $\ell$  and  $\sigma_k^2$  which are referred to as length and amplitude hyperparameters.  $\ell$  determines how close two points need to be to be highly correlated. Larger values of  $\ell$  generates functions with higher correlation within a larger neighbourhood, as seen in Figure 5.2.  $\sigma_k^2$  does not impact the correlation between  $x$  and  $x'$ , but scales the correlation matrix. In other words, larger  $\sigma_k^2$  increase the size but not rate of fluctuations. This can be seen comparing Figure 5.2a to Figure 5.2b.

Other kernels exist and are used in the literature. Here we briefly discuss when  $k(x, x')$  is a valid kernel. We need the formal notions of symmetry and positive semi-definite.

**Definition 5.9** (Symmetric). A (square) matrix  $\mathbf{A}$  is symmetric if  $\mathbf{A} = \mathbf{A}^T$

**Definition 5.10** (Positive Semi-Definite). An  $n \times n$  matrix  $\mathbf{A}$  is positive semi-definite if  $\mathbf{v}^T \mathbf{A} \mathbf{v} \geq 0$  for all  $\mathbf{v} \in \mathbb{R}^n$ .

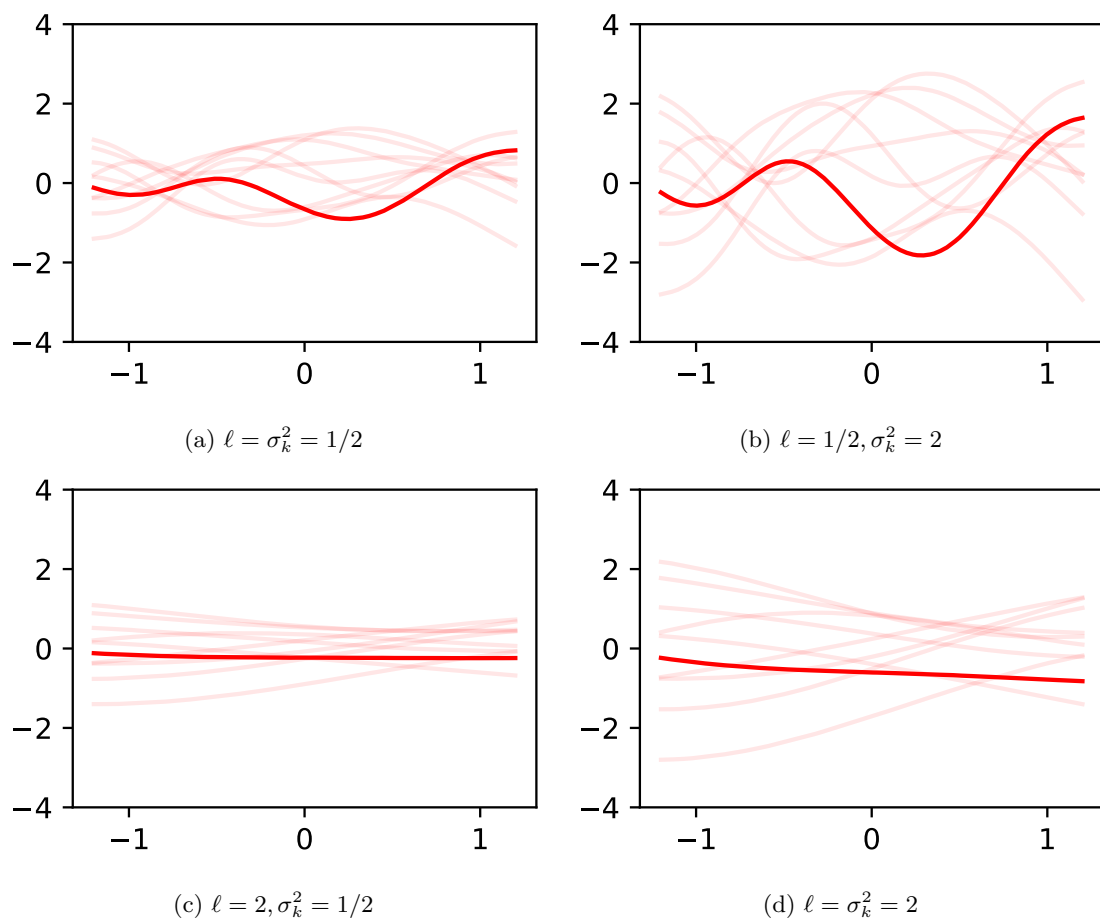


Figure 5.2: Ten realisations of zero mean Gaussian processes with the squared exponential kernel, varying the length and amplitude parameters. The samples were generated using the same seed

**Theorem 5.11.** *Any kernel  $k$  is admissible if the gram matrix*

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_n, x_1) & \dots & \dots & k(x_n, x_n) \end{bmatrix}$$

*associated with  $k$  is symmetric and positive semi-definite for all choices of  $(x_1, \dots, x_n)$ , for finite  $n$*

**Theorem 5.12.** *The Matérn kernel and squared exponential kernels are admissible.*

*Proof.* The proof of the above two theorems are beyond the scope of this thesis, and involves analysis of spectral densities. See Rasmussen and Williams 2008, chapter 4 for more details.  $\square$

## 5.2 Gaussian Process Regression

Given a realisation of a Gaussian process  $f$ , observed at the set of indices  $\mathbf{x}^*$ , we can make inferences on unobserved indices  $\mathbf{x}$  by considering the Gaussian process conditioned on  $f(\mathbf{x}^*)$ . Since  $f$  is a realisation of a Gaussian process, the distribution of  $f(\mathbf{x})|f(\mathbf{x}^*)$  reduces to linear algebra and has a multivariate normal distribution.

**Theorem 5.13** (Conditional Multivariate Normal Distribution is Multivariate Normal). *If*

$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{bmatrix}, \begin{bmatrix} K & K^* \\ (K^*)^T & K^{**} \end{bmatrix} \right),$$

*then*

$$f(\mathbf{x})|f(\mathbf{x}^*) \sim \text{MVN} \left( m(\mathbf{x}) + K^*(K^{**})^{-1}(f(\mathbf{x}^*) - m(\mathbf{x}^*)), K - K^*(K^{**})^{-1}(K^*)^T \right).$$

*Proof.* Since marginal distribution of the multivariate normal distribution, is also multivariate normal,  $f(\mathbf{x}^*) \sim \text{MVN}(m(\mathbf{x}^*), K)$ . Let the inverse of  $\begin{bmatrix} K & K^* \\ (K^*)^T & K^{**} \end{bmatrix}$  be defined as

$$\begin{bmatrix} \tilde{K} & \tilde{K}^* \\ (\tilde{K}^*)^T & \tilde{K}^{**} \end{bmatrix} = \begin{bmatrix} (K - K^*(K^{**})^{-1}(K^*)^T)^{-1} & -(K - K^*(K^{**})^{-1}(K^*)^T)^{-1}K^*(K^{**})^{-1} \\ -(K^{**})^{-1}(K^*)^T(K - K^*(K^{**})^{-1}(K^*)^T)^{-1} & (K^{**})^{-1} + (K^{**})^{-1}(K^*)^T(K - K^*(K^{**})^{-1}(K^*)^T)^{-1}K^*(K^{**})^{-1} \end{bmatrix}$$



by the inverse of a block matrix. Therefore

$$\begin{aligned}
p(f(\mathbf{x})|f(\mathbf{x}^*)) &= \frac{p(f(\mathbf{x}), f(\mathbf{x}^*))}{p(f(\mathbf{x}^*))} \\
&\propto \frac{\exp \left[ -\frac{1}{2} \left( \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{bmatrix} \right)^T \begin{bmatrix} K & K^* \\ (K^*)^T & K \end{bmatrix}^{-1} \left( \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{bmatrix} \right) \right]}{\exp \left[ -\frac{1}{2} (f(\mathbf{x}^*) - m(\mathbf{x}^*))^T (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right]} \\
&= \exp \left[ -\frac{1}{2} \left( \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{bmatrix} \right)^T \begin{bmatrix} K & K^* \\ (K^*)^T & K^{**} \end{bmatrix}^{-1} \left( \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{bmatrix} \right) \right. \\
&\quad \left. + \frac{1}{2} (f(\mathbf{x}^*) - m(\mathbf{x}^*))^T (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right] \\
&= \exp \left[ -\frac{1}{2} \left( (f(\mathbf{x}) - m(\mathbf{x}))^T \tilde{K} (f(\mathbf{x}) - m(\mathbf{x})) \right. \right. \\
&\quad \left. \left. + 2(f(\mathbf{x}) - m(\mathbf{x}))^T \tilde{K}^* (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right. \right. \\
&\quad \left. \left. + (f(\mathbf{x}^*) - m(\mathbf{x}^*))^T \tilde{K}^{**} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right) \right. \\
&\quad \left. + \frac{1}{2} (f(\mathbf{x}^*) - m(\mathbf{x}^*))^T (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right] \\
&\propto \exp \left[ -\frac{1}{2} (f(\mathbf{x}) - m(\mathbf{x}))^T \tilde{K} (f(\mathbf{x}) - m(\mathbf{x})) \right. \\
&\quad \left. - (f(\mathbf{x}) - m(\mathbf{x}))^T \tilde{K}^* (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \right].
\end{aligned}$$

(by removing the terms independent of  $f(\mathbf{x})$ )

Since

$$p(\mathbf{z}) \propto \exp \left( -\frac{1}{2} \mathbf{z}^T \Sigma^{-1} \mathbf{z} + \mathbf{z}^T \mathbf{c} \right) \implies \mathbf{z} \sim \text{MVN}(\Sigma \mathbf{c}, \Sigma),$$

$f(\mathbf{x}) - m(\mathbf{x})|f(\mathbf{x}^*)$  is multivariate normal with mean

$$\begin{aligned}
-\tilde{K}^{-1} \tilde{K}^* (f(\mathbf{x}^*) - m(\mathbf{x}^*)) &= (K - K^* (K^{**})^{-1} (K^*)^T) \\
&\quad \times (K - K^* (K^{**})^{-1} (K^*)^T)^{-1} K^* (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) \\
&= K^* (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*))
\end{aligned}$$

and covariance matrix

$$\tilde{K}^{-1} = K - K^* (K^{**})^{-1} (K^*)^T$$

by the alternative parameterisation of the multivariate normal distribution as a member of the exponential family of distributions (see Wikipedia contributors 2024, Table of Distributions). Finally, by the linearity of the multivariate normal mean,

$$f(\mathbf{x})|f(\mathbf{x}^*) \sim \text{MVN} \left( m(\mathbf{x}) + K^* (K^{**})^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)), K - K^* (K^{**})^{-1} (K^*)^T \right).$$

□

We can use this to fit a Gaussian process to set of indices  $\mathbf{x}^*$  with observations  $f(\mathbf{x}^*)$ . This can be used in an iterative process where  $f(x)$  may be expensive to compute and by treating  $f$  as a Gaussian process realisation, the function can be probabilistically interpolated for unobserved

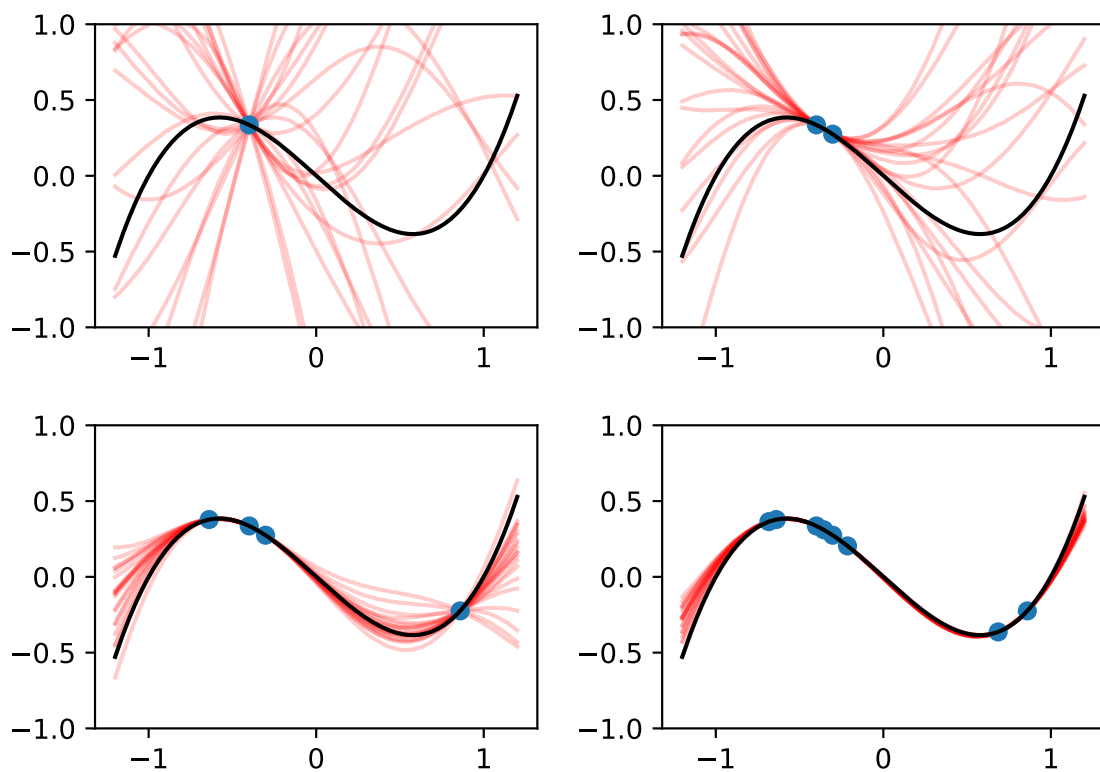


Figure 5.3: Sequence of Gaussian process regressions on the target function (black)  $f(x) = x(x-1)(x+1)$ , after 1, 2, 4, and 8 observations in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was zero mean and had a squared exponential kernel. The hyperparameters were fixed with  $\ell = 2.7$  and  $\sigma_k^2 = 1.1$

$f(x)$ . Figure 5.3 empirically demonstrates that as the number of points that the Gaussian process is conditioned on increases, the variance in the sample paths decreases.

### Observation Variance

For most functions, model outputs, or processes desirable for approximating through Gaussian process regression, it may not be possible to observe  $f(\mathbf{x})$  directly, but observations may be noisy. The simplest assumption is that the observations are of the form

$$f_o(\mathbf{x}^*) = f(\mathbf{x}^*) + \varepsilon$$

where  $\varepsilon \sim \text{MVN}(\mathbf{0}, \sigma_o^2 I)$ . Under these assumptions,  $\text{Cov}(f_o(\mathbf{x}^*), f_o(\mathbf{x}^*)) = K^{**} + \sigma_o^2 I$ , where  $K^{**} = \text{Cov}(f(\mathbf{x}^*), f(\mathbf{x}^*))$  matrix of  $f(\mathbf{x}^*)$  without noise. Therefore the conditional distribution of our unobserved function outputs given noisy observations

$$f(\mathbf{x})|f_o(\mathbf{x}^*) \sim \text{MVN}\left(m(\mathbf{x}) + K^*(K^{**} + \sigma_o^2 I)^{-1}(f(\mathbf{x}^*) - m(\mathbf{x}^*)), K - K^*(K^{**} + \sigma_o^2 I)^{-1}(K^*)^T\right).$$

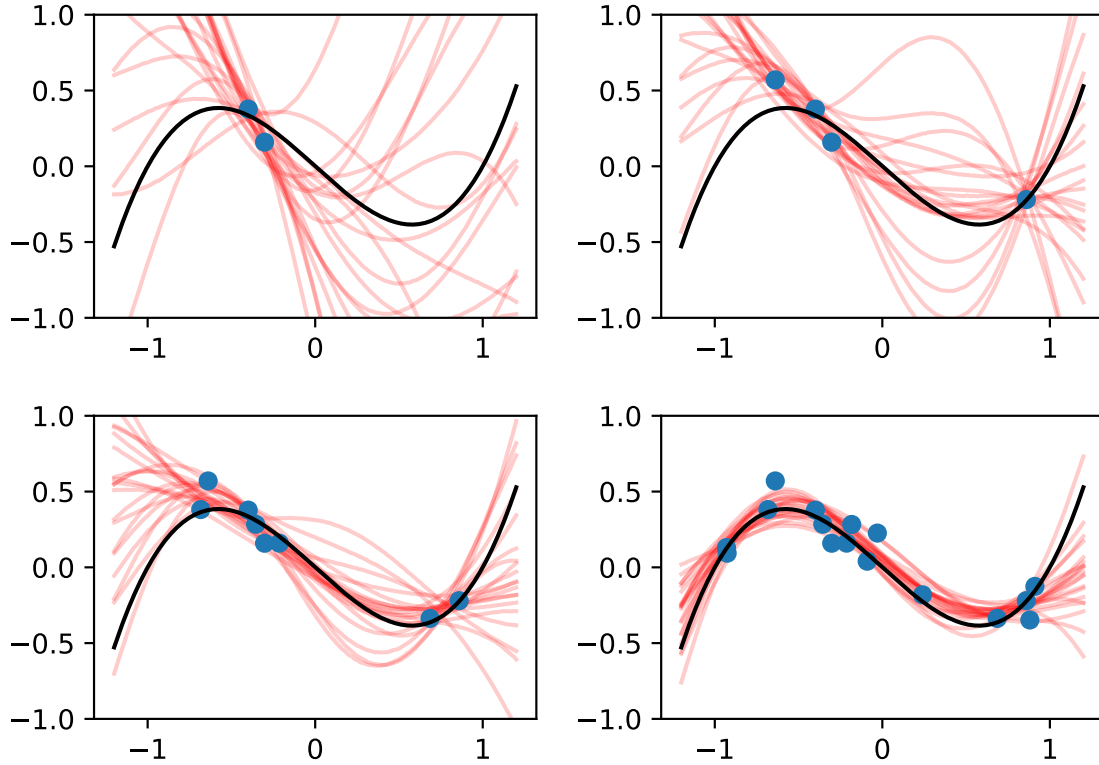


Figure 5.4: Sequence of Gaussian process regressions on the target function (black)  $f(x) = x(x - 1)(x + 1)$ , after 2, 4, 8, and 16 observations of  $f(x_i) + \varepsilon_i$ , where  $\varepsilon_i$  is i.i.d.  $\text{MVN}(0, \sigma_o^2)$  with  $\sigma_o^2 = 0.01$  in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was 0 mean and had a squared exponential kernel. The hyperparameters were fixed with  $\ell = 2.7$  and  $\sigma_k^2 = 1.1$

The observations  $f_o(\mathbf{x}^*)$  contain less information than when  $f(\mathbf{x}^*)$  is directly observed, and hence interpolating to  $\mathbf{x}$  or even  $\mathbf{x}^*$  naturally has a greater degree of uncertainty as seen in Figure 5.4.

### 5.3 Model Selection

After deciding to use a Gaussian process to approximate a function, there is still a variety of choices that need to be made, particularly regarding the covariance kernel. The first choice needs to be the class of kernel used.

#### Kernel

The appropriate choice of kernel will depend on the properties of the target function  $f$  to be regressed to. In the case of estimating an extremely stochastic distribution (such as the price of a stock over time), a kernel with a high degree of mean square differentiability would be inappropriate. On the other hand a Matérn 1/2 kernel may be appropriate, since it is not mean square differentiable. If it is known that the target function is smooth, such as a polynomial function or  $\sin(x)$ , then the choice of squared exponential kernel is the most appropriate kernel.

#### Hyperparameters

The kernel hyperparameters  $\ell$  and  $\sigma_k^2$  are generally not fixed *a priori*. Similarly, the observation variance  $\sigma_o^2$  may not be known. There are two main (frequentist) ways to fit these hyperparameters: maximum likelihood estimation, and leave-one-out cross validation.

Defining the likelihood  $\mathcal{L}(\ell, \sigma_k^2, \sigma_o^2) := p(f(\mathbf{x}^*) | \ell, \sigma_k^2, \sigma_o^2)$  in the usual way, the maximum likelihood estimates are

$$\{\hat{\ell}, \hat{\sigma}_k^2, \hat{\sigma}_o^2\} := \arg \max_{\{\ell, \sigma_k^2, \sigma_o^2\}} \mathcal{L}(\ell, \sigma_k^2, \sigma_o^2) \quad (5.1)$$

which is equivalent to minimising

$$-\ln(\mathcal{L}) = \frac{1}{2} [\ln(|K^{**}(\ell, \sigma_k^2) + \sigma_o^2|) + (f(\mathbf{x}^*) - m(\mathbf{x}^*))^T (K^{**}(\ell, \sigma_k^2) + \sigma_o^2)^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*)) + c].$$

The covariance matrix generated by the choice of kernel  $K^{**}$  is explicitly written with its dependence on  $\ell$  and  $\sigma_k^2$ . Here  $c$  is a constant.

Leave-one-out cross validation aims to maximise the predictive log probability:

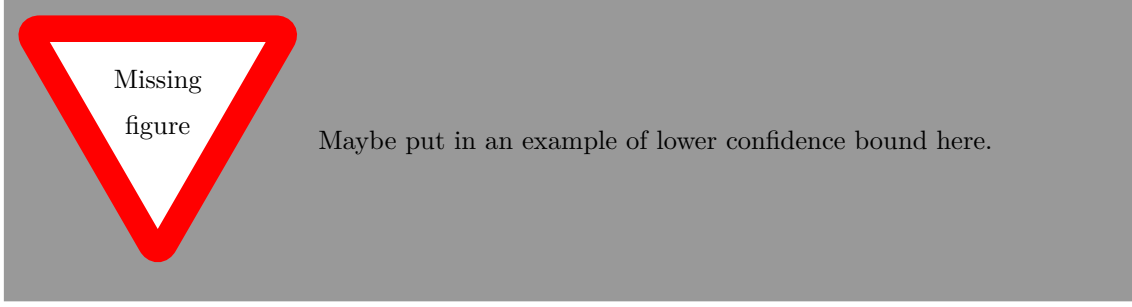
$$\{\tilde{\ell}, \tilde{\sigma}_k^2, \tilde{\sigma}_o^2\} := \arg \max_{\ell, \sigma_k^2, \sigma_o^2} \sum_i \ln p(f_i(\mathbf{x}^*) | f_{-i}(\mathbf{x}^*), \ell, \sigma_k^2, \sigma_o^2), \quad (5.2)$$

where  $f_i(\mathbf{x}^*) | f_{-i}(\mathbf{x}^*)$  is the distribution of the  $i$ th element of  $f(\mathbf{x}^*)$  conditioned on the rest of the observed data excluding that element (represented by  $f_{-i}(\mathbf{x}^*)$ ).  $f_i(\mathbf{x}^*) | f_{-i}(\mathbf{x}^*)$  can be found by Theorem 5.13. Computationally efficient methods for calculating the predictive log probability (Equation 5.2) that avoid having to invert the covariance matrix for every summand element exist. In particular it can be shown that  $f_i(\mathbf{x}^*) | f_{-i}(\mathbf{x}^*)$  has mean

$$f_i(\mathbf{x}^*) - m_i(\mathbf{x}^*) - [(K^{**} + \sigma_o^2 I)^{-1} (f(\mathbf{x}^*) - m(\mathbf{x}^*))]_i / [(K^{**} + \sigma_o^2 I)^{-1}]_{ii}$$

and variance  $1/[(K^{**} + \sigma_o^2 I)^{-1}]_{ii}$ , where both the mean and covariance are (surprisingly) independent of  $f_i(\mathbf{x}^*)$  (Rasmussen and Williams 2008, p. 116).

Recent work has shown that at least under specific conditions, the leave-one-out estimates for the scale hyperparameter are more robust to a larger family of target functions (Naslidnyk et al.



2024), and the broader literature seems to favor leave-one-out cross validation (for example see Gutmann and Cor 2016).

Finally there is scope for a Bayesian approach to model selection. By setting priors on the hyperparameters (sometimes called hyper-priors) and using the likelihood as described in the maximum likelihood estimation approach (Equation 5.1), a posterior distribution can be contrived. Samples could then be taken from the posterior density of the hyperparameters. Alternatively a point estimate could be drawn by choosing the maximum a posteriori estimate (i.e. the mode of the posterior density). Posterior samples of the hyperparameters, can then capture some of the model fit uncertainty, unlike in a point estimate.

## 5.4 Bayesian Acquisition Functions

Gaussian processes may be a useful approximation of  $\mathbb{E}[\mathcal{D}(\theta)]$ . We can express the Gaussian process surrogate model as  $\mathcal{D}_{GP}(\theta)$ , trained on samples of  $\mathcal{D}(\theta)$ . Considering the approximate Bayesian computation described in Algorithm 7, samples are only accepted when  $\mathcal{D}(\theta)$  is small. Therefore we care most about accurately approximating  $\mathcal{D}(\theta)$  where  $\mathbb{E}[\mathcal{D}(\theta)]$  is small since the probability of acceptance elsewhere is negligible. Therefore we focus our model sampling where we predict the Gaussian process is small, or where the variance of the Gaussian process is large (hence the true values are highly uncertain), to avoid unnecessary model runs that may be extremely time consuming.

These ideas are formalised by Bayesian acquisition functions  $\mathcal{A}(\theta)$  which describe the desirability of sampling from  $\theta$  as a combination of low posterior mean and uncertainty.

### Lower Confidence Bound

The lower confidence bound acquisition function is a calculation of the lower bound of the confidence interval of the regressed Gaussian process. It is a function of both the posterior mean and variance, weighted according to a constant  $\eta$ .

**Definition 5.14** (Lower Confidence Bound). *The lower confidence bound of a Gaussian process  $f$  at  $x$  given some observations  $\mathbf{x}^*$  is*

$$\mathcal{A}_{\text{LCB}}(x) := \mathbb{E}[f(x)|f(\mathbf{x}^*)] - \eta\sqrt{\text{Var}[f(x)|f(\mathbf{x}^*)]}$$

For example, when  $\eta = 1.96$ ,  $\mathcal{A}_{\text{LCB}}(x)$  returns the lower bound of the 95% confidence interval at  $x$ . In problems where a global minimum is to be estimated, and where  $f$  is regressed on realisations of a model, the next point to sample from the model would then be chosen  $\arg \min_x \mathcal{A}_{\text{LCB}}(x)$ . Larger

$\eta$  will prioritise exploration of the space of  $x$ , whereas small  $\eta$  will continue to sample around areas of confirmed low mean.

$\eta$  can also be replaced by  $\eta(t)$ , where  $t$  is the number of points that have been regressed on. Generally  $\eta(t)$  is chosen to be an increasing function, so that exploration is given more weight over time. Some theoretical results regarding optimal choice of  $\eta(t)$  are given in Srinivas et al. 2010, and are highly dependent on choice of covariance function and dimensionality of the parameter space.

## Probability of Improvement

The probability of improvement is simply a measure of how probable it is that an observation at  $x$  is better than the previous best observation.

**Definition 5.15.** *The probability of improvement of a Gaussian process  $f$  at  $x$  given some observations  $\mathbf{x}^*$  is*

$$\mathcal{A}_{\text{PI}}(x) := \Pr(f(x) < \mu^*)$$

where  $\mu^* := \min_{x^* \in \mathbf{x}^*} f(x^*)$ .

Unlike the lower confidence bound, we choose  $\arg \max_x \mathcal{A}_{\text{PI}}(x)$ , as the point which is most likely to be better than our current best.

The probability of improvement can also be expressed as

$$\mathcal{A}_{\text{PI}}(x) = \Pr(\min(f(x) - \mu^*, 0) < 0),$$

which motivates the form of the next acquisition function.

## Expected Improvement

A similar acquisition function to the probability of improvement is the expected improvement function. Rather than returning a the probability that  $f(x)$  is better (lower) than the current best, it also takes into account how large that improvement is likely to be.

**Definition 5.16.** *The expected improvement of a Gaussian process  $f$  at  $x$  given some observations  $\mathbf{x}^*$  is*

$$\mathcal{A}_{\text{EI}}(x) := \mathbb{E}[\min(f(x) - \mu^*, 0)],$$

where  $\mu^* := \min_{x^* \in \mathbf{x}^*} f(x^*)$ .

The next point to be sampled from  $\arg \min_x \mathcal{A}_{\text{EI}}(x)$  is the point where we expect  $\mu^*$  to have the largest improvement, if it is indeed improved.<sup>1</sup> Both the probability of improvement and expected improvement do not require a choice of hyperparameter such as  $\eta$ , but more exploration can be encouraged by slightly altering the probability of improvement to  $\Pr(f(x) < \mu^* + \epsilon)$ , and the expected improvement to  $\mathbb{E}[\min(f(x) - (\mu^* + \epsilon), 0)]$ .  $\epsilon$  allows for new samples to be  $\epsilon$  worse (larger) than the current best sample. This is beneficial in the case where finding the exact global minimum may not be the target, but rather exploring areas close to the minimum.

---

<sup>1</sup>Bayesian acquisition function are conventionally employed to find the maximum of an unknown function, and so generally the expected improvement is maximised. However in the context of  $\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta})$ , we want to find the minimum. Therefore we have reframed the expected improvement as a function to be minimised.

## Part II

# Calibrating Parameters for a *P.* *vivax* Model





# Chapter 6

## Methods

We now develop a method for calibrating model parameters for a *P. vivax* model. In order to validate the method we create a simulation study. This is where we simulate our observed data from the model, and then use that data to recover the parameters used in creating our observed data.

### 6.1 Creation of Synthetic Data

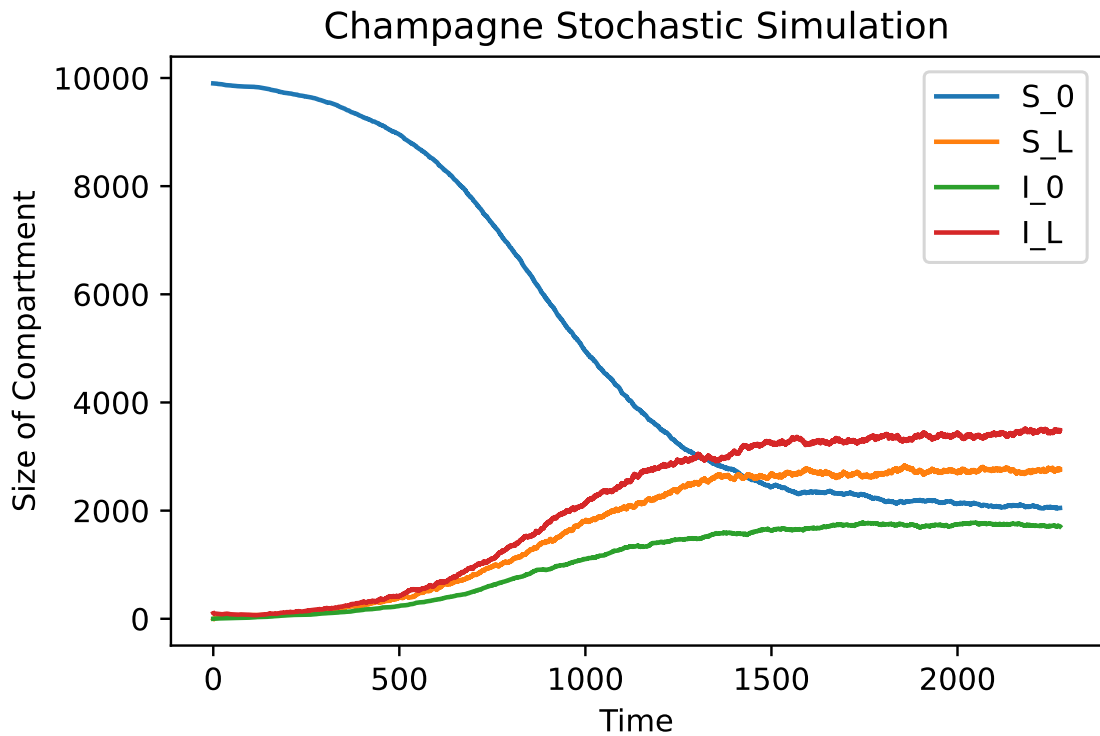


Figure 6.1: A Doob-Gillespie Simulation of the model described by Champagne et al. 2022 with  $\alpha = 0.4$ ,  $\beta = 0.4$ ,  $\gamma_L = 1/223$ ,  $\lambda = 0.04$ ,  $f = 1/72$ ,  $r = 1/60$ , and  $\delta = 0$ . The population was 10000, with 100 initial infections (both blood and liver stage  $I_L$ ).

We investigated the model by Champagne et al. 2022 as described in Section 3.1. A malaria

Table 6.1: The parameters used to simulate a *P. vivax* outbreak using the model described by Champagne et al. 2022

Parameter description	Parameter	Value	Units
effective blood stage treatment proportion	$\alpha$	0.1235	None
effective liver stage treatment proportion	$\beta$	0.429	None
rate of liver stage disease clearance	$\gamma_L$	1/383	1/days
rate of infection	$\lambda$	0.01	1/days
rate of relapse	$f$	1/69	1/days
rate of blood stage disease clearance	$r$	1/60	1/days.

epidemic was simulated using the Doob-Gillespie algorithm shown in Figure 6.1, using a population size of 10,000, and initial infected population of 100 (with both liver and blood stage infection). The parameters used closely followed those reported in Champagne et al. 2022, with the exact parameters used reported in Table 6.1. To simplify our analysis, we have assumed that case importation is negligible, and therefore we have set  $\delta = 0$ . From initialisation, the simulation was run for 200,000 time steps, after which, the model was assumed to have reached steady state behaviour. The number of time steps was chosen as the stopping criteria, because the time the model takes to reach a steady state is highly dependent on the scales of the parameters. Time steps adapt with the scale of the parameters.

Table 6.2: Observed synthetic data  $\mathbf{y}^{\text{obs}} := \{\iota_{\text{obs}}, \pi_{\text{obs}}, i_{\text{obs}}, p_{\text{obs}}\}$  from the simulation in Figure 6.1.

Parameter Description	Parameter	Observed Value
Weekly incidence at epidemic steady state	$\iota_{\text{obs}}$	461
Prevalence at epidemic steady state	$\pi_{\text{obs}}$	5205
Incidence in the first month of the epidemic	$i_{\text{obs}}$	42
Prevalence after one month of the epidemic	$p_{\text{obs}}$	87

Within our assumed model framework, new infections which instantly undergo radical cure do not change the size of each compartment. These infections should contribute to our incidence, even though they are not calculated. To account for this, silent infections, we calculated the number of additional infections to be Poisson distribution with rate  $\Delta t \times \alpha\beta\lambda(I_L + I_0)S_0/N$ , where  $\Delta t$  is the time between events (one time step). The (simulated) observed data was taken to be from the simulated case counts (incidence) and prevalence (as the absolute number of people infected) of the simulated epidemic, described in Table 6.2.

## 6.2 Model Simulations and Discrepancy Function

New epidemics were simulated as above, with 200,000 events and at least 30 days (to allow for calculation of incidence in the first month of the epidemic), with parameters  $\boldsymbol{\theta} = \{\alpha, \beta, \gamma_L, \lambda, f, r\}$ . For each model  $y(\boldsymbol{\theta}) = \{\iota, \pi, i, p\}$  was calculated with the same method as  $\mathbf{y}^{\text{obs}}$ , where the interpretation of each parameter is described in Table 6.2.

We defined the discrepancy function to be  $L_2$  norm of the relative differences

$$\mathcal{D}(\boldsymbol{\theta}) = \mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r) := \sqrt{\left(\frac{\iota - \iota_{\text{obs}}}{\iota_{\text{obs}}}\right)^2 + \left(\frac{\pi - \pi_{\text{obs}}}{\pi_{\text{obs}}}\right)^2 + \left(\frac{i - i_{\text{obs}}}{i_{\text{obs}}}\right)^2 + \left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2}.$$

Relative difference was chosen to limit the impact between the scale differences of the summary

statistics.

### 6.3 Gaussian Process and Initialisation

We approximated  $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})]$  with a Gaussian process  $d_{\mathcal{GP}}(\boldsymbol{\theta})$  surrogate model.  $d_{\mathcal{GP}}(\boldsymbol{\theta})$  was regressed on sample means

$$\overline{\ln \mathcal{D}(\boldsymbol{\theta})} := \frac{1}{30} \sum_{j=1}^{30} \ln \mathcal{D}_j(\boldsymbol{\theta})$$

where the  $\ln \mathcal{D}_j(\boldsymbol{\theta})$ s are i.i.d. samples generated by model runs. Each evaluation of the discrepancy function was run in parallel using a supercomputer. The samples  $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$  were assumed to be noisy observations of  $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})] + \varepsilon$ , where  $\varepsilon \sim N(0, \sigma_o^2)$ , by the central limit theorem.  $\sigma_o^2$  was assumed to be independent of  $\boldsymbol{\theta}$ , and has the natural interpretation as the variance of the sample mean.  $d_{\mathcal{GP}}(\boldsymbol{\theta})$  was assumed to have unknown constant mean  $m_{\mathcal{GP}}$ , and kernel

$$k(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i) = \sigma_k^2 \left(1 + z_i + \frac{z_i^2}{3}\right) \exp(-z_i)$$

where

$$z_i = \sqrt{5 \sum_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \left( \frac{\theta_i - \theta'_i}{\ell_{\boldsymbol{\theta}}} \right)^2}.$$

This kernel is, a Matérn kernel with  $\nu = 5/2$  and automatic relevance determination - i.e. each parameter  $\theta \in \boldsymbol{\theta}$  was scaled by  $\ell_{\boldsymbol{\theta}}$ . In effect, this assigns each parameter its own length hyperparameter. It is important that each parameter has its own length scale because each parameter has different scales, and has varying degrees of impact on the mean of the log discrepancy.

Table 6.3: Conservative upper bounds for parameters to be calibrated. Values were informed by Champagne et al. 2022; White et al. 2016. All lower bounds were zero.

Parameter	Upper Bound	Unit
Proportion of treatment clearing blood stage disease $\alpha$	1	
Proportion of treatment clearing liver stage disease $\beta$	1	
Rate of liver stage disease clearance $\gamma_L$	1/30	1/days
Rate of infection $\lambda$	1/10	1/days
Rate of relapse $f$	1/14	1/days
Rate of blood stage disease clearance $r$	1/14	1/days

All parameters to be calibrated were given conservative upper bounds after considering values reported in the literature.  $d_{\mathcal{GP}}(\boldsymbol{\theta})$  was fit over this compact subspace of the whole parameter space.

Latin hypercube sampling was used to initialise 50 samples of the parameter space (scaled to be between zero and the upper bounds described in Table 6.3). For each set of parameters,  $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$  was generated. The hyperparameters described in Table 6.4 were optimised using leave one out cross validation of the log predictive likelihood described in Equation 5.2, and  $d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta})$  was fit to the samples.

Table 6.4: Hyperparameters used in training  $d_{\mathcal{GP}}(\boldsymbol{\theta})$ .

Hyperparameter	Description
$\sigma_k^2$	Matérn kernel amplitude
$\sigma_o^2$	Observation variance ( $\text{var}(\ln \mathcal{D}(\boldsymbol{\theta}))$ )
$\ell_\alpha$	Length parameter associate with $\alpha$
$\ell_\beta$	Length parameter associate with $\beta$
$\ell_{\gamma_L}$	Length parameter associate with $\gamma_L$
$\ell_\lambda$	Length parameter associate with $\lambda$
$\ell_f$	Length parameter associate with $f$
$\ell_r$	Length parameter associate with $r$
$m_{\mathcal{GP}}$	Gaussian process mean

---

**Algorithm 8** Gaussian process approximation of  $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$  with Bayesian updating

---

**Input:** Initial values for  $\boldsymbol{\theta}$ , lower and upper bounds for  $\boldsymbol{\theta}$ , initial Gaussian process model  $d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta})$   
**Output:** Synthetic likelihood  $\hat{L}(\boldsymbol{\theta})$

```

for  $t = 1$  to 500 do
   $\boldsymbol{\theta}^{(t)} \leftarrow \arg \min_{\boldsymbol{\theta}} \mathcal{A}_{\text{EI}}(\boldsymbol{\theta})$ 
  Sample  $\ln \mathcal{D}(\boldsymbol{\theta}^{(t)})$ 
  if  $t \leq 6$  then ▷ Once per parameter
     $j \leftarrow t$ 
    Create  $\mathbf{s}_j$ , 15 evenly spaced values from 0 to the upper bound of  $\theta_j$  in Table 6.3
    for  $k$  in 1 to 15 do
       $\theta_j^{(t)} \leftarrow s_{jk}$ 
      Sample  $\ln \mathcal{D}(\boldsymbol{\theta}^{(t)})$ 
    end for
  else
     $j \leftarrow t \bmod 6$  ▷ Iterating over  $\boldsymbol{\theta}$ 
    for 4 repeats do
      Sample  $U_j \sim \text{Unif}(0, m_j)$ , with  $m_j$  being  $\theta_j$ 's upper bound
       $\theta_j^{(t)} \leftarrow U_j$ 
      Sample  $\ln \mathcal{D}(\boldsymbol{\theta}^{(t)})$ 
    end for
  end if
  if  $t \bmod 50 == 0$  then ▷ Every 50 iterations
    Reoptimise the Gaussian process hyperparameters using leave-one-out cross validation
  end if
  Update  $d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$  on the new samples
end for
return  $d_{\mathcal{GP}}^{(500)}(\boldsymbol{\theta})$ 

```

---

## 6.4 Bayesian Acquisition and Parameter Updates

The Gaussian process was optimised over 500 iterations. Each iteration involved minimising the expected improvement and obtaining a new sample  $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$ . For each iteration, one of the variables  $\theta_j \in \boldsymbol{\theta}$  was chosen, and multiple samples of  $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$  were taken using multiple values for  $\theta_j$ . This was in order to improve the values for the length parameters, but also to explore the parameter space more widely. This step is highly parallelisable. The hyperparameters were reoptimised every 50 iterations, as the number of samples increased. The full procedure is specified in Algorithm 8.

Given a the set of previously sampled parameters  $\boldsymbol{\theta}^*$ , and  $\mu_* := \min_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} \mathbb{E}(d_{\mathcal{GP}}(\boldsymbol{\theta}_*))$ , the expected improvement with exploration

$$\mathcal{A}_{\text{EI}}(\boldsymbol{\theta}) := \mathbb{E}[\min(d_{\mathcal{GP}}(\boldsymbol{\theta}) - (\mu^* + 0.1), 0)],$$

was minimised using a gradient descent algorithm. The gradient descent was initialised at  $\boldsymbol{\theta}^\dagger$ , where  $\boldsymbol{\theta}^\dagger$  was a combination of the best parameters yet observed  $\boldsymbol{\theta}^{**} := \arg \min_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} d_{\mathcal{GP}}(\boldsymbol{\theta}^*)$ , and values between 0 and the upper bounds described in 6.3 distributed uniformly at random. Each  $\theta_j^\dagger$  was set independently, with

$$\Pr(\theta_j^\dagger = \theta_j^{**}) = 1/2 = \Pr(\theta_j^\dagger \text{ uniformly distributed}).$$

We could have chosen random initialisation but we do not expect it to improve the algorithms success, because for samples where  $\mathcal{A}_{\text{EI}}(\boldsymbol{\theta})$  is very small, particularly if  $d_{\mathcal{GP}}(\boldsymbol{\theta}^*)$  is large, the gradient of  $\mathcal{A}$  can be negligible, causing the convergence criteria to be met prematurely. Furthermore starting only at the current minimum will increase the possibility of being stuck in local minima.

The final fitted Gaussian process  $d_{\mathcal{GP}}^{(500)}(\boldsymbol{\theta})$  is an approximation of  $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ . Since the variance of the sample mean  $\text{var}(\overline{\ln \mathcal{D}(\boldsymbol{\theta})})$  is estimated by  $\sigma_o^2$ , the variance of the the log discrepancy function  $\ln \mathcal{D}(\boldsymbol{\theta})$  is approximately  $30\sigma_o^2$ . We then used moment matching assuming that  $\mathcal{D}(\boldsymbol{\theta})$  is approximately log-normally distributed distribution, and hence can be approximated by

$$\hat{\mathcal{D}}(\boldsymbol{\theta}) \sim \text{LN} \left( \mathbb{E}[d_{\mathcal{GP}}^{(500)}(\boldsymbol{\theta})], 30\sigma^2 \right).$$

Therefore using approximate Bayesian computation described in Algorithm 7, the probability of sampling and accepting a  $\boldsymbol{\theta}$  is  $\Pr(\boldsymbol{\theta}) \Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta})$ , where  $L(\boldsymbol{\theta}) := \Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta})$  approximates the true likelihood  $\mathcal{L}(\boldsymbol{\theta})$ . Finally, we substitute in our approximation  $\hat{\mathcal{D}}$  for  $\mathcal{D}$ , to create our synthetic likelihood

$$\hat{L}(\boldsymbol{\theta}) := \Pr(\hat{\mathcal{D}}(\boldsymbol{\theta}) < \epsilon | \boldsymbol{\theta}). \quad (6.1)$$

Since

$$\ln \hat{\mathcal{D}}(\boldsymbol{\theta}) \sim N \left( \mathbb{E}[d_{\mathcal{GP}}^{(500)}(\boldsymbol{\theta})], 30\sigma^2 \right),$$

we can express  $\hat{L}$  as

$$\hat{L}(\boldsymbol{\theta}) = \Pr(\ln \hat{\mathcal{D}}(\boldsymbol{\theta}) < \ln \epsilon | \boldsymbol{\theta}).$$

This final step is the crux of our methodology. By recovering a likelihood the standard frequentist and Bayesian methods for parameter estimation are once again usable.

The Gaussian process and Gaussian process regression was implemented using TensorFlow

(Martín Abadi et al. 2015), and all code is available at [https://github.com/jaycrick/masters\\_project](https://github.com/jaycrick/masters_project).

# Chapter 7

## Results

We outline the results of the method first by validating that each step of the method had plausible outputs, and then by reporting the estimated parameters based on the observed data.

### 7.1 Validation

Table 7.1: Final Gaussian process hyperparameters

Hyperparameter	Final value
$\sigma_o^2$	0.07
$\sigma_k^2$	0.707
$\ell_\alpha$	0.324
$\ell_\beta$	0.715
$\ell_{\gamma_L}$	0.010
$\ell_\lambda$	0.006
$\ell_f$	0.016
$\ell_r$	0.016
$m_{\mathcal{GP}}$	0.879

The final hyperparameters, are reported in Table 7.1. To ensure that our expected improvement and hyperparameter optimisation functions converged, we plotted the initial trainings in Figures 7.2 and 7.3. The plots have a smooth curve, demonstrating that both algorithms converged at a good rate.

The violin plot of discrepancy function in Figure 7.1 is well dispersed suggesting a good amount of exploration of the parameter space has been done. A large number of samples were in areas of low mean log discrepancy, and so this suggests a good exploration exploitation trade-off.

Figure 7.4 suggests that after 400 iterations  $d_{\mathcal{GP}}(\boldsymbol{\theta})$  fits to  $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$  well. The  $\lambda$  slice has the sharpest minimum, whereas  $\beta$  does not appear to have much impact on the discrepancy. As the number of iterations increased, the Gaussian process visibly improved at fitting to the mean. Interim iterations of  $d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$  can be seen in the appendix.

### 7.2 Parameter estimation

For parameter estimation, we typically look at likelihood instead of discrepancy. This is done by inverting that distribution function (see Equation 6.1). In Table 7.2 we present the global

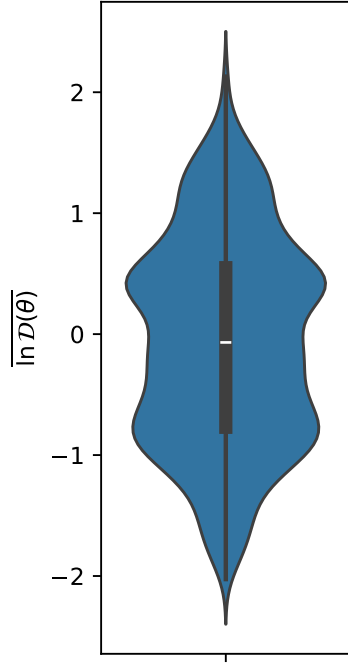
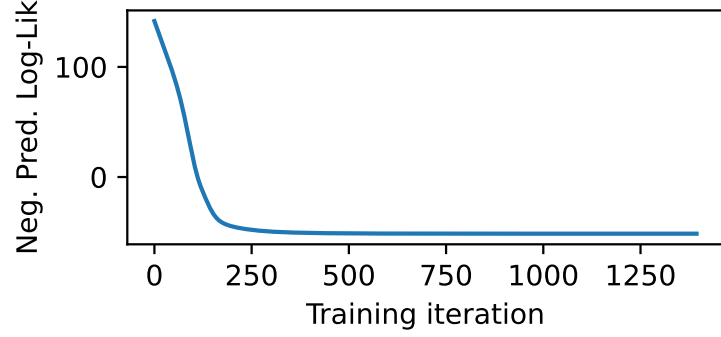
Figure 7.1:  $\ln \mathcal{D}(\theta)$  violin plot

Figure 7.2: Hyperparameter training

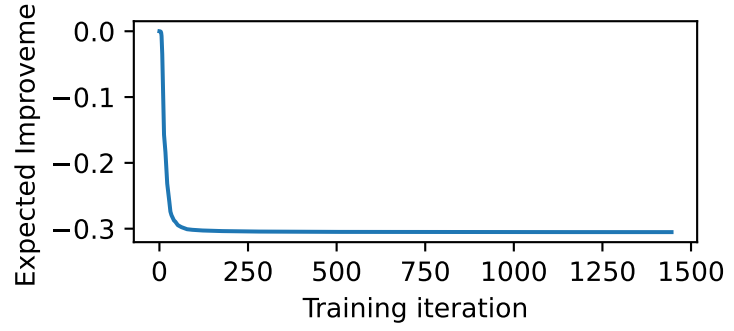
Figure 7.3: Finding  $\arg \min_{\theta} \mathcal{A}_{\text{EI}}(\theta)$ 

Table 7.2: Estimates of our model parameters. The maximum likelihood estimate (MLE) of the true parameters using  $\hat{L}$ . The maximum slice estimate was the one-dimensional maximum likelihood estimate where all other parameters are held constant at the true value.

Parameter	True	MLE	ML Slice Estimate
$\alpha$	0.124	0.153	0.02
$\beta$	0.429	0.555	0.63
$\gamma_L$	0.0026	0.006	0.005
$\lambda$	0.01	0.01	0.01
$f$	0.014	0.024	0.017
$r$	0.017	0.023	0.022



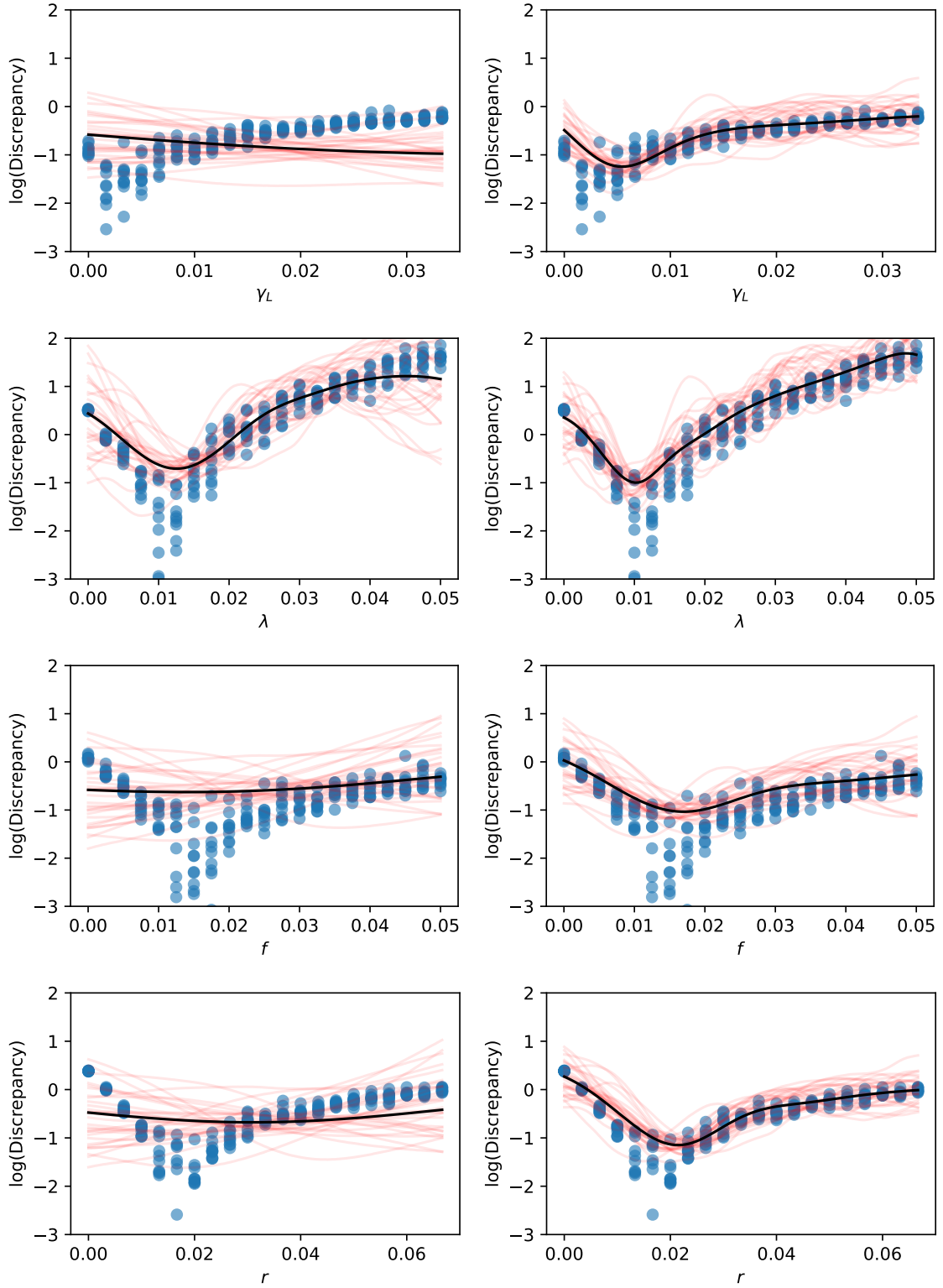


Figure 7.4: The left column of figures is the Gaussian process after initialisation  $d_{\mathcal{GP}}^{(0)}(\theta)$ . The black line is  $\mathbb{E}(d_{\mathcal{GP}}^{(0)}(\theta))$ , and the red lines are multiple realisations of  $d_{\mathcal{GP}}^{(0)}(\theta)$ . The right column of figures is after 400 sampling iterations, with the black line being  $\mathbb{E}(d_{\mathcal{GP}}^{(400)}(\theta))$ . The blue dots are realisations of  $\ln \mathcal{D}(\theta)$ .  $d_{\mathcal{GP}}$  has not been trained on these realisations. The parameters are varied univariately, with all other parameters fixed at the true parameters.

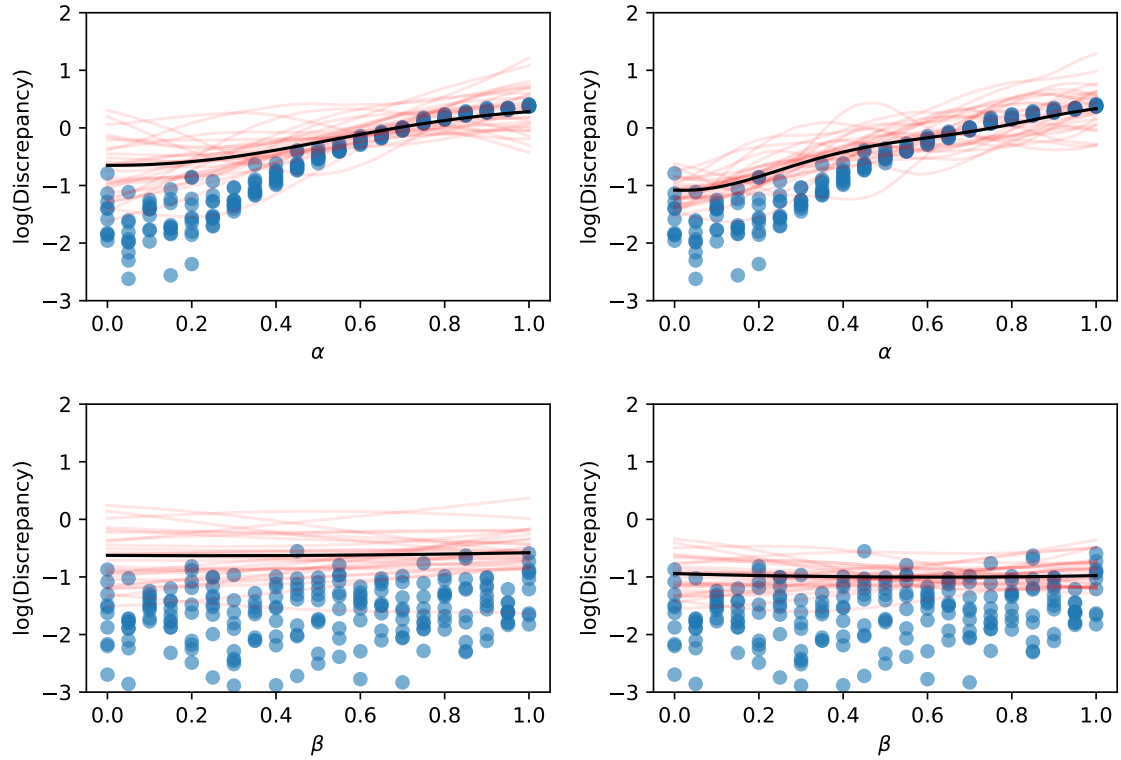


Figure 7.5: Gaussian process approximations of the treatment parameters, as with Figure 7.4

maximum likelihood estimate found by finding the maximum synthetic likelihood  $\arg \max_{\theta} \hat{L}(\theta)$ .

In Figure 7.6 we show our final likelihood function estimate from the trained Gaussian process. For simplicity we only present the likelihood across parameter slices, where the parameter of interest is allowed to vary, while all other parameters are held constant at their true values. The maximum likelihood values for each parameter slice is also presented in Table 7.2, but do not vary much from the maximum likelihood estimates, and the true values. For this reason, we can be confident that our synthetic likelihood is a good approximation of the true likelihood, despite the true likelihood being infeasible to compare to. Certain parameters such as  $\lambda$ ,  $r$ , and  $\gamma_L$  have sharper likelihood peaks than parameters such as  $\alpha$  and  $\beta$ , suggesting the data we used may lend itself to estimating parameters such as  $\lambda$  over the treatment parameters, although all likelihood slices calculated were unimodal.

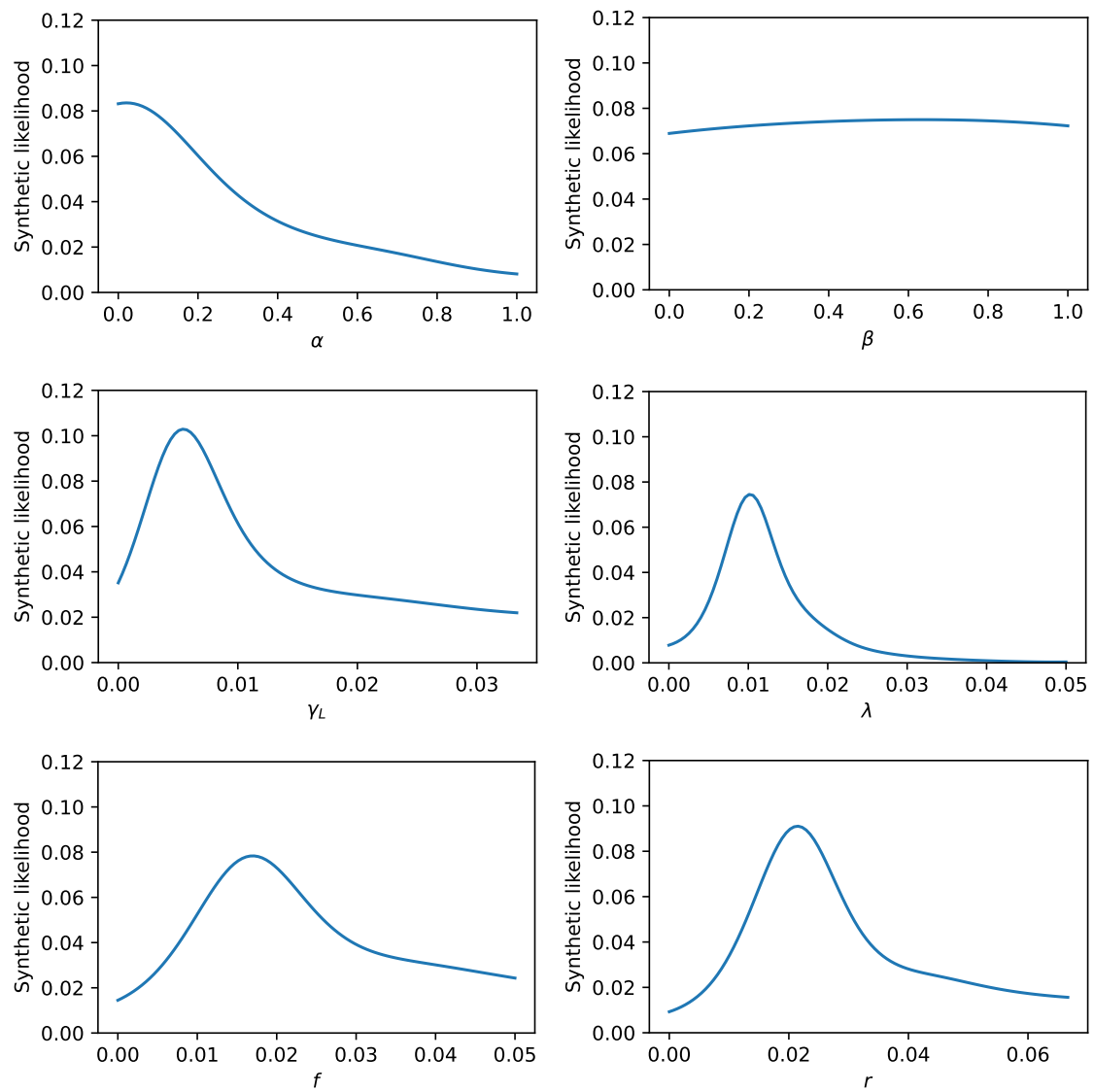


Figure 7.6: Final univariate synthetic likelihoods  $\hat{L}(\theta)$  after 400 sampling iterations. All values not shown were fixed at the true parameters.



## Chapter 8

# Discussion

To our knowledge, the use of the synthetic likelihood as described above has not been used to calibrate a malaria model before. Champagne et al. calibrate the model we use by finding the ODE equilibrium, and fitting a single parameter  $\lambda$  to incidence data. Note that this comes with limitations. For example, this relies on data being observed when the disease is roughly at equilibria in the population. This is not desirable for modelling outbreaks or seasonality effects. Champagne et al.'s approach parameter calibration does not work in these scenarios.

In comparison, not only were we able to effectively recover the true  $\lambda$  from a synthetic run, we were able to recover all model parameters simultaneously.

This methodology is very robust to both frequentist and Bayesian inference. Under a Bayesian framework, since we can evaluate the synthetic likelihood for any  $\theta$ , we can use a Metropolis Hasting sampler, to obtain samples from a distribution approximately equal to the posterior distribution  $\Pr(\theta|\mathbf{y}^{\text{obs}})$ . Alternatively, standard frequentist inference can also be used on our synthetic likelihood  $\hat{L}$ . For example, numerically approximating the observed Fisher information matrix

$$F(\hat{\theta}) = -\frac{\partial \ln \hat{L}}{\partial \theta \partial \theta^T}(\hat{\theta})$$

allows us to do hypothesis testing and construct confidence intervals, since asymptotically  $\hat{\theta} \sim N(\theta, F^{-1}(\hat{\theta}))$  (see Fahrmeir, Hennevogel, and Tutz 2013).

Although the likelihood free procedure we outlined for this model closely resembles Gutmann and Cor 2016, there are a few significant changes that improve on the method outlined in that manuscript.

The most significant change is the choice to model the sample mean as a noisy Gaussian process, rather than modelling the discrepancy function as a noisy Gaussian process. Changing this assumption means that we are not subject to the distributional assumption of normal noise and constant variance with respect to  $\mathcal{D}(\theta)$ . Once we have an approximation for the mean as a function of  $\theta$ , we could then choose a distribution that scales with the mean. This is analogous to the generalised linear modelling framework. For example, it may be reasonable to assume that  $\mathcal{D}(\theta)$  follows a Gamma distribution. Letting  $\mu(\theta) := \mathbb{E}[\mathcal{D}(\theta)]$ , we can approximate  $\mathcal{D}(\theta)$  with  $\hat{\mathcal{D}}(\theta) \sim \text{Gamma}\left(\frac{\mu(\theta)}{\phi}, \frac{1}{\phi}\right)$ , where  $\frac{\mu(\theta)}{\phi}, \frac{1}{\phi}$  are the shape and rate parameters. Trivially

$$\mathbb{E}[\hat{\mathcal{D}}(\theta)] = \frac{\mu(\theta)}{\phi} / \frac{1}{\phi} = \mu(\theta),$$

and for a fixed  $\phi$ ,  $\text{var}[\mathcal{D}(\boldsymbol{\theta})] = \phi\mu(\boldsymbol{\theta})$ , so the variance scales linearly with the mean. If this behaviour is observed empirically then such a choice will be preferable, since then  $\hat{L}(\boldsymbol{\theta})$  will be a better approximation of the true likelihood. This can be done with any single parameter distribution with fixed variance structure.

Alternatively the sample variance could also be modelled with a different Gaussian process  $s_{\mathcal{GP}}^2(\boldsymbol{\theta})$ . Any two parameter distribution could be moment matched by the two Gaussian processes to get a more accurate  $\hat{L}(\boldsymbol{\theta})$ . Therefore if empirically we observe that  $\mathcal{D}(\boldsymbol{\theta})$  is approximately Gamma distributed, then we could approximate  $\mathcal{D}(\boldsymbol{\theta})$  with

$$\hat{\mathcal{D}}(\boldsymbol{\theta}) \sim \text{Gamma}\left(\frac{\mu^2(\boldsymbol{\theta})}{\sigma^2(\boldsymbol{\theta})}, \frac{\mu(\boldsymbol{\theta})}{\sigma^2(\boldsymbol{\theta})}\right),$$

where  $\text{var}[\hat{\mathcal{D}}(\boldsymbol{\theta})] = \sigma^2(\boldsymbol{\theta})$  and  $\mathbb{E}(\hat{\mathcal{D}}(\boldsymbol{\theta})) = \mu(\boldsymbol{\theta})$  as required.

The mean and variance do not have a linked structure in the discrepancy function which we used for the Champagne model. This can be seen particularly in the  $\lambda$  slice in Figure 7.4. For  $\boldsymbol{\theta}$  with  $\lambda < 0.03$ ,  $\text{var}(\ln \mathcal{D}(\boldsymbol{\theta}))$ , is small, and  $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta})) \approx 0$ . However for  $\lambda \approx 0.02$ , we also have  $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})] \approx 0$ , but the variance is observably larger. Therefore if we had modelled the sample variance of our log discrepancy as a noisy Gaussian process  $s_{\mathcal{GP}}^2(\boldsymbol{\theta})$  then we could have approximated  $\mathcal{D}(\boldsymbol{\theta})$  with  $\hat{\mathcal{D}}(\boldsymbol{\theta}) \sim \text{LN}(\mathbb{E}(d_{\mathcal{GP}}(\boldsymbol{\theta})), \sigma^2(\boldsymbol{\theta}))$ , where  $\mu(\boldsymbol{\theta}) := \mathbb{E}(d_{\mathcal{GP}}(\boldsymbol{\theta}))$  and  $\sigma^2(\boldsymbol{\theta}) := \mathbb{E}(s_{\mathcal{GP}}^2(\boldsymbol{\theta}))$ .

Empirically it is not surprising that the variance is not constant across the parameter space, or even mean dependent. Disease model behaviour is heavily dependent on the values of the parameters. For example around bifurcation points a slight change in parameters may lead to a disease model that dies out some of the time, but reaches equilibrium in other runs. Here we would expect that  $\text{var}[\mathcal{D}(\boldsymbol{\theta})]$  to be large. This threshold is important in disease modelling, and is called  $R_0$ : the expected number of secondary cases a single infectious individual causes in a completely susceptible population. But when  $\boldsymbol{\theta}$  is changed only a small amount such that  $R_0 < 1$ , and so the disease consistently dies out, the  $\text{var}[\mathcal{D}(\boldsymbol{\theta})]$  will be close to 0. Since the model run will always end with a disease free population and the summary statistics such as incidence or prevalence will always be 0. This is likely what is happen for very small  $\lambda$  in Figure 7.4.

This highlights another problem. Around bifurcation points such as near where  $R_0 = 1$ , it is expected that  $\mathbb{E}(\mathcal{D}(\boldsymbol{\theta}))$  behaves erratically. Gutmann and Cor use a squared exponential kernel for their Gaussian process approximation which cannot capture this behaviour without making any length scales very large. On the other hand, a Matérn 5/2 kernel as we used is still smooth, but allows enough flexibility that sharp changes in the target function do not require overcompensation in the length scales. For an example, demonstrating the utility of the Matérn kernel over the squared exponential in the case of a non smooth function, see Jones 2021. Another possible solution is to use a Student- $t$  process to approximate the discrepancy, as it has heavier tails, so is more forgiving to sudden jumps. The multivariate Student- $t$  distribution has some properties analogous to the multivariate normal distribution. This includes an analytic solution to the conditional distribution, similar to Theorem 5.13. For more details see Shah, Wilson, and Ghahramani 2014.

Gutmann and Cor 2016 use the lower confidence bound acquisition function, where the explo-

ration parameter is the slowly increasing function

$$\eta_t := \sqrt{2 \ln \left( \frac{t^{2/d+2} \pi^2}{3\varepsilon} \right)}.$$

This is chosen because under the exponentiated quadratic kernel, and compact support, the lower confidence bound samples are shown to be no regret with high probability. There are multiple issues with this. The first is that Gutmann and Cor seem to have inherited this form from Brochu, Cora, and Freitas 2010. However the citation in Brochu, Cora, and Freitas 2010 wrongly reproduces the result in Srinivas et al. 2010,<sup>1</sup> which should be

$$\eta_t := \sqrt{2 \ln \left( \frac{t^{2d+2} \pi^2}{3\varepsilon} \right)}.$$

When we tried both of these exploration parameters, the choice of  $\varepsilon$  between  $(0, 1)$  largely lead to repeated sampling from the same set of parameters, even for very small  $\varepsilon$ . This is similar to the behaviour reported in Gutmann and Cor 2016. Secondly, Gelman et al. do not restrict the parameter space to a compact subset of the space, and so theoretical guarantees of no regret are not valid. Finally, Srinivas et al. find this assuming zero mean Gaussian processes. Gutmann and Cor assume a quadratic mean prior.

Even though we do consider a compact subset of the parameter space, we did not use a zero mean Gaussian process, and we used the Matérn kernel, therefore we did not want to use this formulation of the lower confidence bound, and instead chose expected improvement.

The quadratic mean assumption in Gutmann and Cor 2016 is also problematic. If the mean of the Gaussian process is trained on a set of data which concave near a local minimum, no matter which acquisition function is chosen, areas away from the local minimum may not be sampled from, since the mean function will dominate the predicted behaviour of  $\mathcal{D}(\theta)$ , and so exploration will be minimal. This also may explain the behaviour of the acquisition function sampling close to the same point repeatedly.

Although we have validated this method on a relatively low dimensional  $\theta$ , for models with high dimensionality, it is likely that construction of a synthetic likelihood will have greater comparative benefits to other methods such as approximate Bayesian computation. This is because as the dimensionality increases, the curse of dimensionality means that randomly drawn points will be increasingly further away on average, and so many more samples from the prior distribution function are likely to produce  $\mathcal{D}(\theta) > \epsilon$ , particularly if  $\mathcal{D}(\theta)$  is small in a small region. Optimising the acquisition function each time encourages efficient sampling from areas that are likely to be beneficial to sample from.

One way to reduce the computational overhead of this method would be to reduce the number of samples were used to calculate the sample mean. This paper used 30 to ensure convergence to the normal distribution, however less could be taken with the trade off of a larger observation variance. The sample mean can be calculated in parallel, so the rate limiting step is minimising the acquisition function each iteration. Rather than sampling uniformly across a single parameter, multivariate noise could be added to  $\theta$  to sample multiple sample means at once. Resources should be maximally allocated to calculate as many  $\mathcal{D}(\theta)$  as feasible in one step, split between

---

<sup>1</sup>One Python package that implements the methodology in Gutmann and Cor 2016 notes this error, see: <https://github.com/elfi-dev/elfi/blob/dev/elfi/methods/bo/acquisition.py>

multiple repeats for the sample mean, and multiple  $\theta$ s for better training of the Gaussian process approximation.



## Chapter 9

# Conclusion

Malaria, a significant global health challenge, continues to burden the global health system. Mathematical disease models are increasingly being harnessed to alleviate this burden. This thesis explores ways to maximise the impact of epidemiological models by ensuring they accurately simulate public health scenarios, with a specific focus on the complex issue of *P. vivax* malaria.

The complicated lifecycle of *P. vivax* poses a tough hurdle in disease modelling, particularly with respect to asymptomatic cases and relapses. This complexity is problematic during the parameter calibration stage, where traditional methods that rely on being able to compute a likelihood are not viable. Likelihood-free methods, while effective, come with a high computational cost due to repeated and inefficient model runs. This has led some researchers to calibrate models using deterministic approximations, failing to capture the stochastic models' uncertainty. There is an obvious need for a better calibration methodology.

This research used Gaussian processes to approximate the distribution of the discrepancy function used in approximate Bayesian computation for any set of parameters. The true likelihood function was approximated by using the Gaussian process to create a synthetic likelihood. Empirical evidence demonstrated that this methodology successfully recovered parameters of a *P. vivax* model given simulated data while mitigating the computational overhead. Additionally, this thesis laid out possible further extensions to the methodology.

In conclusion, this thesis has demonstrated the plausibility of a new, more robust method of calibrating malaria models, particularly for *P. vivax*. The use of Gaussian processes and synthetic likelihoods has proven effective in overcoming the infeasibility of traditional calibration methods and the computational cost of likelihood-free calibration methods. Future research should focus on refining these techniques and exploring their application to other infectious diseases. These advancements are crucial for developing more effective interventions and ultimately to help support achieving malaria eradication.



# Bibliography

- Abramowitz, Milton and Irene A. Stegun, eds. (2013). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover books on mathematics. New York, NY: Dover Publ. 1046 pp. ISBN: 978-0-486-61272-0.
- Acuña-Zegarra, Manuel Adrian et al. (July 2021). “COVID-19 optimal vaccination policies: A modeling study on efficacy, natural and vaccine-induced immunity responses”. In: *Mathematical Biosciences* 337. Publisher: Elsevier, p. 108614. DOI: 10.1016/j.mbs.2021.108614. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8095066/> (visited on 06/19/2024).
- Adams, John H. and Ivo Mueller (Sept. 2017). “The Biology of Plasmodium vivax”. In: *Cold Spring Harbor Perspectives in Medicine* 7.9, a025585. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a025585. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5580510/> (visited on 03/24/2023).
- Aron, Joan L. and Robert M. May (1982). “The population dynamics of malaria”. In: *The Population Dynamics of Infectious Diseases: Theory and Applications*. Ed. by Roy M. Anderson. Boston, MA: Springer US, pp. 139–179. ISBN: 978-1-4899-2901-3. DOI: 10.1007/978-1-4899-2901-3\_5. URL: [https://doi.org/10.1007/978-1-4899-2901-3\\_5](https://doi.org/10.1007/978-1-4899-2901-3_5).
- Brochu, Eric, Vlad M. Cora, and Nando de Freitas (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. arXiv: 1012.2599 [cs.LG].
- Champagne, Clara et al. (Jan. 2022). “Using observed incidence to calibrate the transmission level of a mathematical model for Plasmodium vivax dynamics including case management and importation”. In: *Mathematical Biosciences* 343, p. 108750. ISSN: 00255564. DOI: 10.1016/j.mbs.2021.108750. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0025556421001541> (visited on 08/22/2023).
- Cowman, Alan F. et al. (2016). “Malaria: Biology and Disease”. In: *Cell* 167.3. Type: Review, pp. 610–624. DOI: 10.1016/j.cell.2016.07.055. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994000411&doi=10.1016%2fj.cell.2016.07.055&partnerID=40&md5=81d9b4c51fe738ac66e0c8561b12c5bf>.
- Fahrmeir, Ludwig, W. Hennevogl, and Gerhard Tutz (2013). *Multivariate Statistical Modelling Based on Generalized Linear Models*. OCLC: 1066189579. New York, NY: Springer. ISBN: 978-1-4899-0010-4.
- Gani, Raymond and Steve Leach (Dec. 13, 2001). “Transmission potential of smallpox in contemporary populations”. In: *Nature* 414.6865, pp. 748–751. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/414748a. URL: <https://www.nature.com/articles/414748a> (visited on 06/10/2024).

- Gelman, Andrew et al. (2014). *Bayesian data analysis*. Third edition. Texts in statistical science series. Boca Raton London New York: CRC Press, Taylor and Francis Group. 667 pp. ISBN: 978-1-4398-4095-5.
- Gillespie, Daniel T. (Dec. 1977). “Exact stochastic simulation of coupled chemical reactions”. In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361. ISSN: 0022-3654, 1541-5740. DOI: 10.1021/j100540a008. URL: <https://pubs.acs.org/doi/abs/10.1021/j100540a008> (visited on 06/19/2024).
- (July 22, 2001). “Approximate accelerated stochastic simulation of chemically reacting systems”. In: *The Journal of Chemical Physics* 115.4, pp. 1716–1733. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1378322. URL: <https://pubs.aip.org/jcp/article/115/4/1716/451187/Approximate-accelerated-stochastic-simulation-of> (visited on 06/19/2024).
- Gutmann, Michael U. and Jukka Cor (2016). “Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models”. In: *Journal of Machine Learning Research* 17.125, pp. 1–47. ISSN: 1533-7928. URL: <http://jmlr.org/papers/v17/15-017.html> (visited on 04/28/2024).
- Hagenaars, T. J., C. A. Donnelly, and N. M. Ferguson (Apr. 2006). “Epidemiological analysis of data for scrapie in Great Britain”. en. In: *Epidemiology and Infection* 134.2, pp. 359–367. ISSN: 0950-2688, 1469-4409. DOI: 10.1017/S0950268805004966. URL: [https://www.cambridge.org/core/product/identifier/S0950268805004966/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S0950268805004966/type/journal_article) (visited on 03/26/2024).
- Jones, Andy (July 31, 2021). *The Matérn class of covariance functions*. Andy Jones. URL: <https://andrewcharlesjones.github.io/journal/maternal-kernels.html> (visited on 06/18/2024).
- Keeling, Matthew James and Pejman Rohani (2008). *Modeling infectious diseases in humans and animals*. OCLC: ocn163616681. Princeton: Princeton University Press. 366 pp. ISBN: 978-0-691-11617-4.
- Kerr, Cliff C. et al. (July 26, 2021). “Covasim: An agent-based model of COVID-19 dynamics and interventions”. In: *PLOS Computational Biology* 17.7. Ed. by Manja Marz, e1009149. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009149. URL: <https://dx.plos.org/10.1371/journal.pcbi.1009149> (visited on 06/19/2024).
- Martín Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: <https://www.tensorflow.org/>.
- Milner, Danny A. (Jan. 2018). “Malaria Pathogenesis”. en. In: *Cold Spring Harbor Perspectives in Medicine* 8.1, a025569. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a025569. URL: <http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a025569> (visited on 03/24/2023).
- Naslidnyk, Masha et al. (2024). *Comparing Scale Parameter Estimators for Gaussian Process Interpolation with the Brownian Motion Prior: Leave-One-Out Cross Validation and Maximum Likelihood*. arXiv: 2307.07466 [math.ST].
- Price, R.N. et al. (2020). “Plasmodium vivax in the Era of the Shrinking P. falciparum Map”. English. In: *Trends in Parasitology* 36.6, pp. 560–570. ISSN: 1471-4922. DOI: 10.1016/j.pt.2020.03.009.
- Rasmussen, Carl Edward and Christopher K. I. Williams (2008). *Gaussian processes for machine learning*. 3. print. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press. 248 pp. ISBN: 978-0-262-18253-9.

- Robert, Christian P. and George Casella (2010). *Monte Carlo statistical methods*. 2. ed., softcover reprint of the hardcover 2. ed. 2004. Springer texts in statistics. New York, NY: Springer New York. 645 pp. ISBN: 978-1-4757-4145-2 978-1-4419-1939-7. DOI: 10.1007/978-1-4757-4145-2.
- Shah, Amar, Andrew Gordon Wilson, and Zoubin Ghahramani (2014). *Student-t Processes as Alternatives to Gaussian Processes*. arXiv: 1402.4306.
- Smith, David L. et al. (Apr. 2012). “Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens”. en. In: *PLOS Pathogens* 8.4. Publisher: Public Library of Science, e1002588. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1002588. URL: <https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002588> (visited on 03/28/2023).
- Srinivas, Niranjan et al. (2010). “Gaussian process optimization in the bandit setting: no regret and experimental design”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Haifa, Israel: Omnipress, pp. 1015–1022. ISBN: 9781605589077.
- White, Michael T. et al. (Mar. 30, 2016). “Variation in relapse frequency and the transmission potential of *Plasmodium vivax* malaria”. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1827, p. 20160048. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2016.0048. URL: <https://royalsocietypublishing.org/doi/10.1098/rspb.2016.0048> (visited on 08/22/2023).
- Wikipedia contributors (2024). *Exponential family* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 16-May-2024]. URL: [https://en.wikipedia.org/w/index.php?title=Exponential\\_family&oldid=1202463189](https://en.wikipedia.org/w/index.php?title=Exponential_family&oldid=1202463189).
- World Health Organization (Dec. 2022). *World malaria report 2022*. en. Tech. rep. Geneva: World Health Organization.
- Zekar, Lara and Tariq Sharman (2023). “Plasmodium Falciparum Malaria”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing. URL: <http://www.ncbi.nlm.nih.gov/books/NBK555962/> (visited on 03/24/2023).
- Zha, Wen-ting et al. (2020). “Research about the optimal strategies for prevention and control of varicella outbreak in a school in a central city of China: based on an SEIR dynamic model”. en. In: *Epidemiology and Infection* 148, e56. ISSN: 0950-2688, 1469-4409. DOI: 10.1017/S0950268819002188. URL: [https://www.cambridge.org/core/product/identifier/S0950268819002188/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S0950268819002188/type/journal_article) (visited on 03/26/2024).



# Appendix A

## Additional Theorems and Proofs

**Theorem A.1** (Sums of Independent Poisson Point Processes). *Given independent Poisson point processes  $\{\mathcal{N}_1(t)\}_{t \geq 0}, \{\mathcal{N}_2(t)\}_{t \geq 0}, \dots, \{\mathcal{N}_n(t)\}_{t \geq 0}$ , with intensities  $\lambda_1, \lambda_2, \dots, \lambda_n$ ,*

$$\{\mathcal{N}(t)\}_{t \geq 0} := \{\mathcal{N}_1(t) + \mathcal{N}_2(t) + \dots + \mathcal{N}_n(t)\}_{t \geq 0}$$

*is a Poisson point process with intensity  $\lambda_1 + \lambda_2 + \dots + \lambda_n$ .*

*Proof.* We show that  $\{\mathcal{N}(t) := \{\mathcal{N}_1(t) + \mathcal{N}_2(t) + \dots + \mathcal{N}_n(t)\}_{t \geq 0}\}$  meets each component of Definition 2.1.

1.  $\mathcal{N}(0) := \mathcal{N}_1(0) + \mathcal{N}_2(0) + \dots + \mathcal{N}_n(0) = 0$  since  $\mathcal{N}_i(0) = 0$  by definition of a Poisson point process.
2. We show that  $\mathcal{N}(t_1) - \mathcal{N}(t_0), \mathcal{N}(t_2) - \mathcal{N}(t_1), \dots, \mathcal{N}(t_n) - \mathcal{N}(t_{n-1})$  with  $0 \leq t_0 < t_1 < \dots < t_n$  are independent.

$$\mathcal{N}(t_i) - \mathcal{N}(t_{i-1}) = \underbrace{[\mathcal{N}_1(t_i) - \mathcal{N}_1(t_{i-1})]}_{X_{i1}} + \underbrace{[\mathcal{N}_2(t_i) - \mathcal{N}_2(t_{i-1})]}_{X_{i2}} + \dots + \underbrace{[\mathcal{N}_n(t_i) - \mathcal{N}_n(t_{i-1})]}_{X_{in}}$$

$X_{ik}$  is independent of  $X_{j\ell}$  for  $k \neq \ell$  since  $\mathcal{N}_k$  and  $\mathcal{N}_\ell$  are independent processes.  $X_{ik}$  is independent of  $X_{jk}$  for  $i \neq j$  by the second property of Definition 2.1. Therefore all  $X_{ik}$  are independent of  $X_{j\ell}$  for  $i \neq j$ , and all  $j, k$ . Hence  $\mathcal{N}(t_1) - \mathcal{N}(t_0), \mathcal{N}(t_2) - \mathcal{N}(t_1), \dots, \mathcal{N}(t_n) - \mathcal{N}(t_{n-1})$  with  $0 \leq t_0 < t_1 < \dots < t_n$  are independent.

3. For fixed  $t_1 < t_2$ , and  $i \in \{1, 2, \dots, n\}$ ,

$$\mathcal{N}_i(t_2) - \mathcal{N}_i(t_1) \sim \text{Pois}((t_2 - t_1)\lambda_i).$$

Consider the associated moment generating function of  $\mathcal{N}_i(t_2) - \mathcal{N}_i(t_1)$ ,

$$M_i(z) := \exp(\lambda_i(t_2 - t_1)(\exp(z) - 1)).$$

Therefore the moment generating function of

$$\mathcal{N}(t_2) - \mathcal{N}(t_1) = [\mathcal{N}_1(t_2) - \mathcal{N}_1(t_1)] + [\mathcal{N}_2(t_2) - \mathcal{N}_2(t_1)] + \dots + [\mathcal{N}_n(t_2) - \mathcal{N}_n(t_1)]$$

is

$$M(z) := \prod_{i=1}^n M_i(z) = \exp[(\lambda_1(t_2 - t_1) + \lambda_2(t_2 - t_1) + \cdots + \lambda_n(t_2 - t_1))(\exp(z) - 1)].$$

Therefore  $\mathcal{N}_1(t) + \mathcal{N}_2(t) + \cdots + \mathcal{N}_n(t) \sim \text{Pois}((\lambda_1 + \lambda_2 + \cdots + \lambda_n)t)$  by the uniqueness of the moment generating function. □

**Theorem A.2** (Time to First Event in Poisson Point Process). *Given a Poisson point process  $\{\mathcal{N}(t)\}_{t \geq 0}$  with intensity  $\lambda$ , let  $\tau = \inf\{t | \mathcal{N}(t_0 + t) - \mathcal{N}(t_0) = 1, t > 0\}$ .  $\tau \sim \text{Exp}(\lambda)$  for  $t_0 \geq 0$*

*Proof.*

$$\Pr(\tau > x) = \Pr(\mathcal{N}(t_0 + x) - \mathcal{N}(t_0) = 0) = \frac{(\lambda x)^0 e^{-\lambda x}}{0!} = e^{-\lambda x}$$
□

**Theorem A.3** (Probability of  $i$ th Poisson Process Generating the Next Event). *Consider independent Poisson point processes*

$$\{\mathcal{N}_1(t)\}_{t \geq 0}, \{\mathcal{N}_2(t)\}_{t \geq 0}, \dots, \{\mathcal{N}_n(t)\}_{t \geq 0}$$

*having intensities  $\lambda_1, \lambda_2, \dots, \lambda_n$ . For fixed  $t_0$ , let  $\tau_i := \inf\{t | \mathcal{N}(t_0 + t) - \mathcal{N}(t_0) = 1\}$ . Then*

$$\Pr(\min_i \tau_i = \tau_j) = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}.$$

*Proof.* By Theorem A.2,  $\tau_i \sim \text{Exp}(\lambda_i)$ . Therefore

$$\begin{aligned} \Pr(\min_i \tau_i = \tau_j) &= \int_0^\infty \Pr(\{\tau_i = x\} \cup \bigcup_{j \neq i} \{\tau_j > x\}) dx \\ &= \int_0^\infty \Pr(\{\tau_i = x\} \cup \bigcup_{j \neq i} \{\tau_j > x\}) dx \\ &= \int_0^\infty \Pr(\tau_i = x) \times \prod_{j \neq i} \Pr(\tau_j > x) dx && \text{(by independence)} \\ &= \int_0^\infty \lambda_i \exp(-\lambda_i x) \times \prod_{j \neq i} \exp(-\lambda_j x) dx \\ &= \lambda_i \int_0^\infty \exp(-(\sum_{i=1}^n \lambda_j)x) dx \\ &= \lambda_i \left[ \frac{\exp(-(\sum_{i=1}^n \lambda_j)x)}{\sum_{i=1}^n \lambda_j} \right]_0^\infty \\ &= \frac{\lambda_i}{\sum_{i=1}^n \lambda_j} \end{aligned}$$
□



## Appendix B

# Additional Results

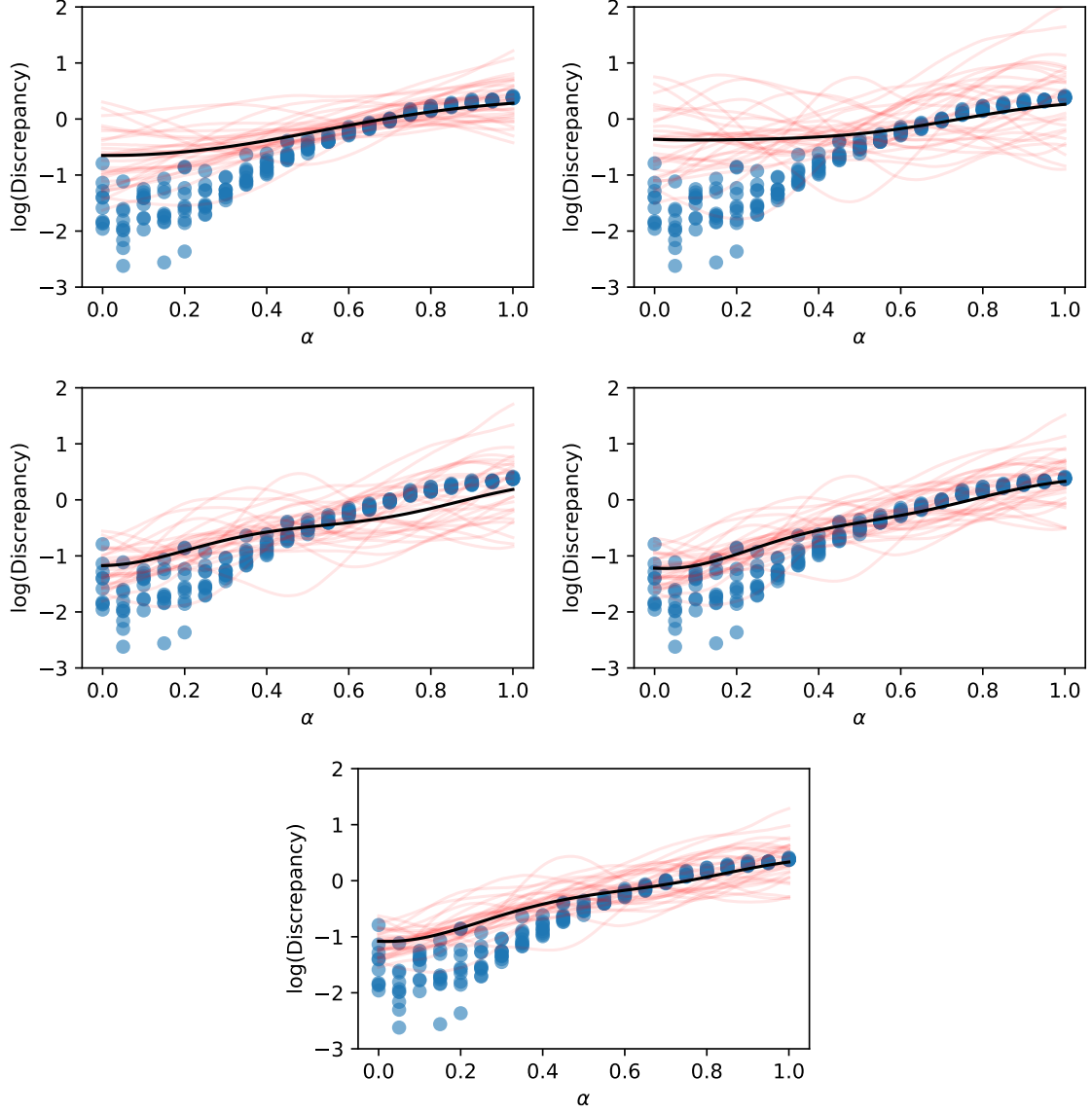


Figure B.1:  $d_{GP}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $\alpha$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .

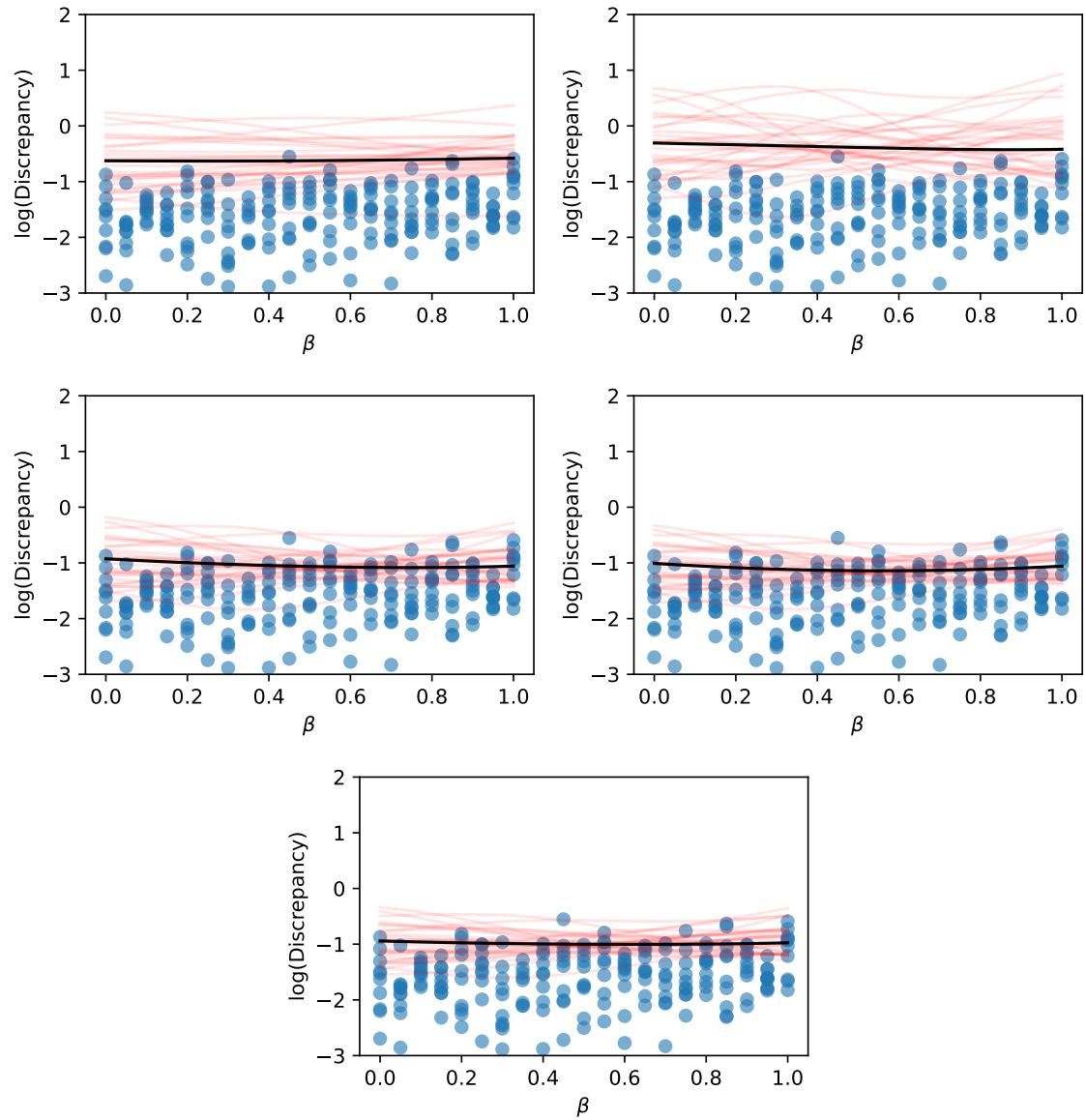


Figure B.2:  $d_{\mathcal{GP}}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $\beta$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .

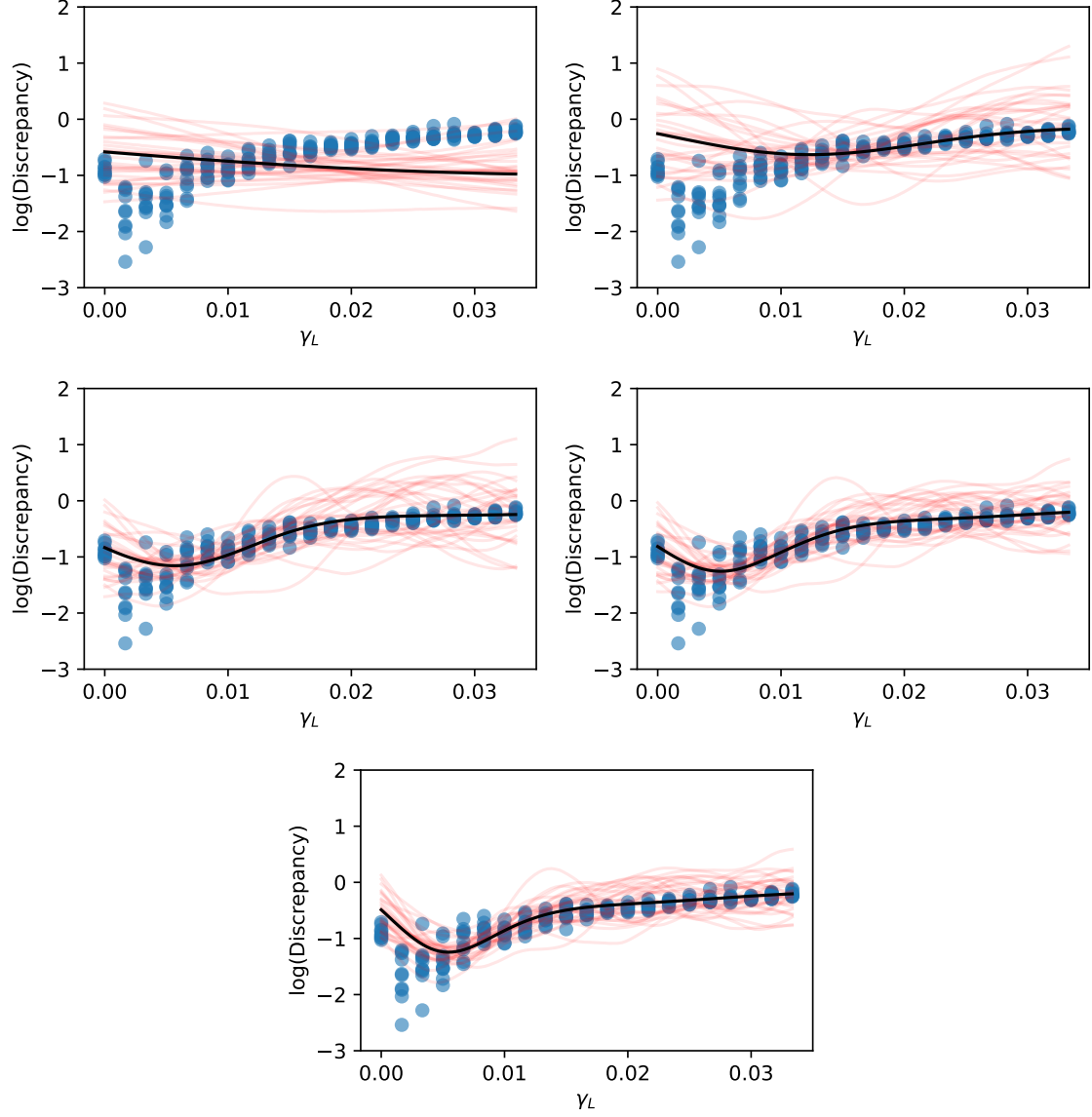


Figure B.3:  $d_{\mathcal{GP}}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $\gamma_L$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .

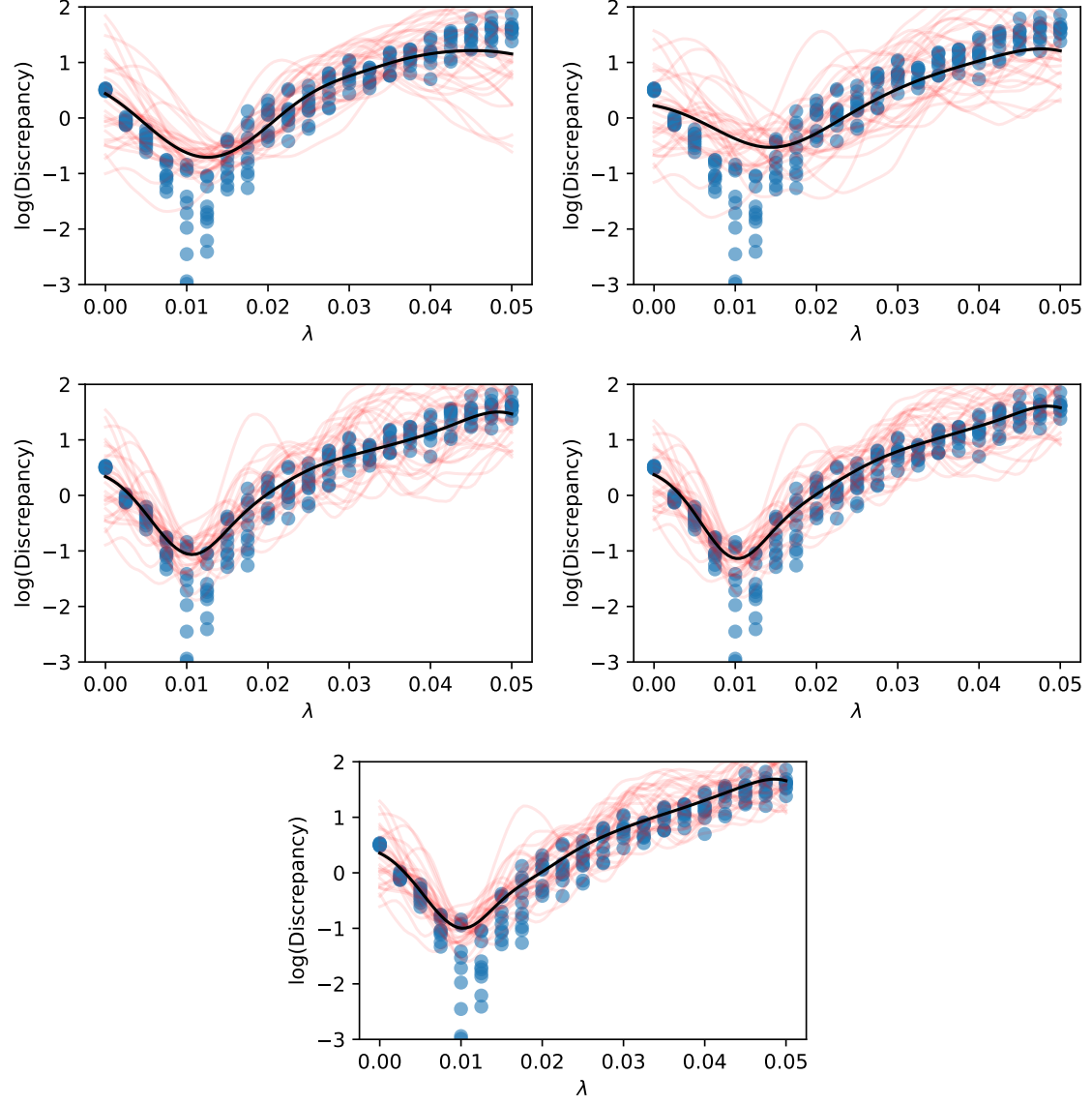


Figure B.4:  $d_{GP}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $\lambda$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .

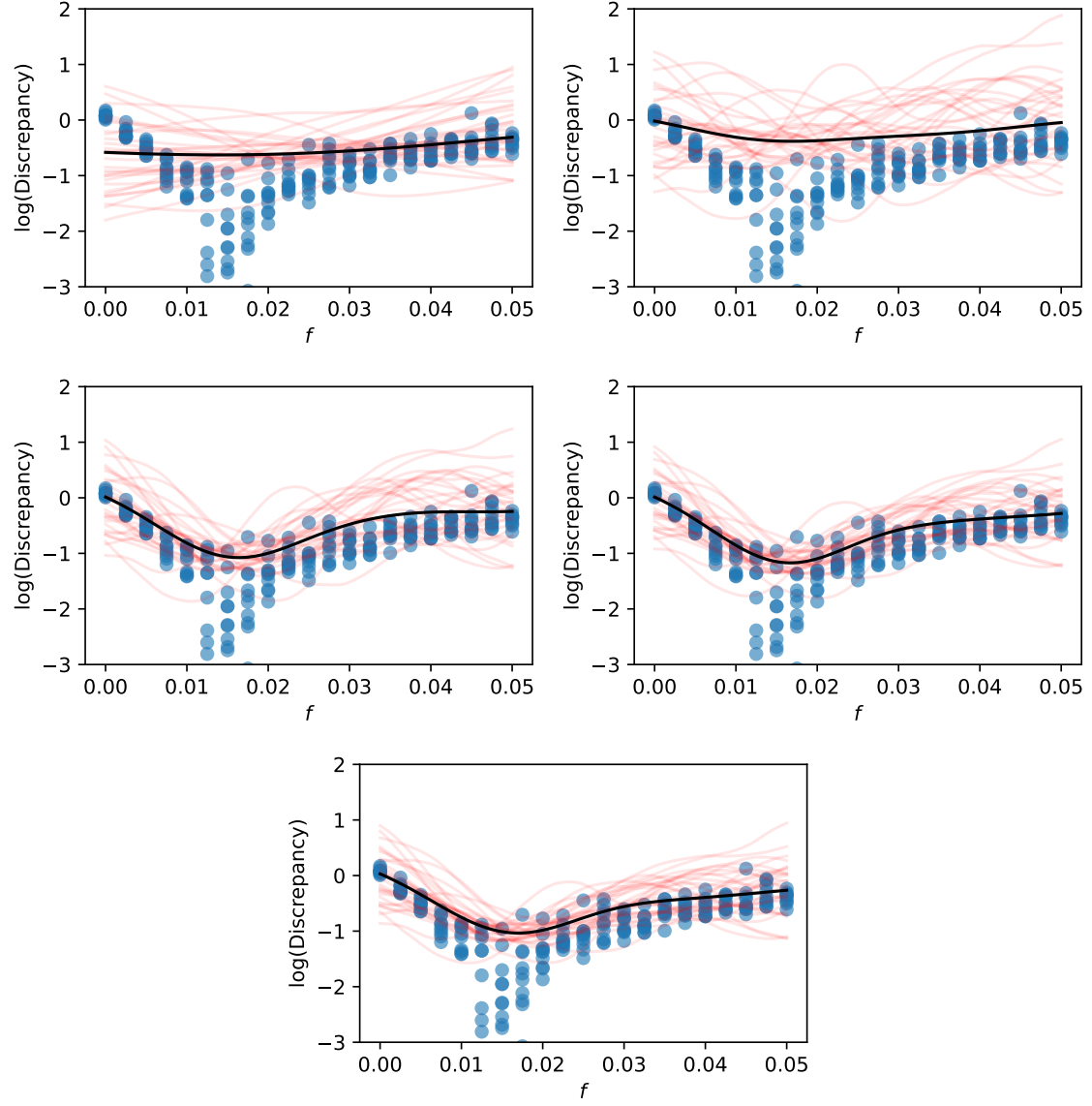


Figure B.5:  $d_{GP}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $f$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .

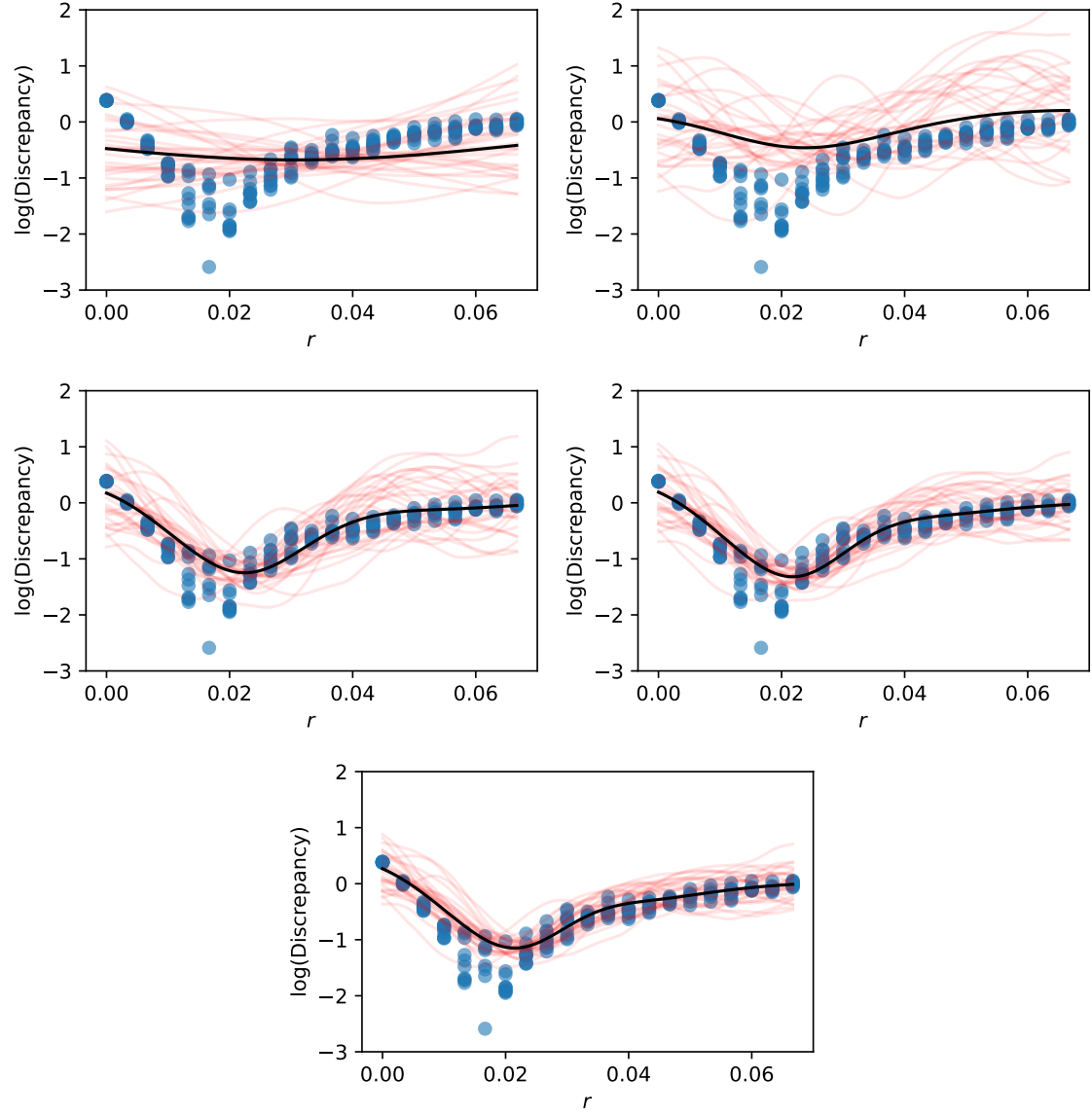


Figure B.6:  $d_{\mathcal{GP}}^{(t)}(\theta)$  approximation of  $\mathbb{E}(\ln \mathcal{D}(\theta))$ , for  $t = 0, 100, 200, 300$ , and  $400$ . Only  $r$  was varied. All other parameters were fixed at the true values. Black line is  $\mathbb{E}(d^{(i)}(\theta))$ . Blue dots are realisations from  $\ln \mathcal{D}(\theta)$ .