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

1 Introduction 1

I Literature Review 3

2 Epidemiological Modelling 5
- 2.1 Deterministic ODE models 6
- 2.2 Stochastic models 8
- 2.3 Doob-Gillespie Algorithm 11
- 2.4 τ-leaping 11

3 Malaria and Malaria Models 13
- 3.1 Malaria 13
- 3.2 Malaria Models 15

4 Parameter Inference 21
- 4.1 Motivation 21
- 4.2 Frequentist Parameter Estimation 21
- 4.3 Bayesian Parameter Estimation 24

5 Gaussian Processes 35
- 5.1 Motivation and Definitions 35
- 5.2 Families of Kernel Function 37

6 Gaussian Process Regression 41
- 6.1 Observation Variance 44
- 6.2 Model Selection 45
- 6.3 Differing mean functions 46
- 6.4 Bayesian Acquisition Functions 46

II Calibrating Parameters for a *P. vivax* Model 49

7 Methods 51
- 7.1 Creation of Synthetic Data 51
- 7.2 Model Simulations and Discrepancy Function 52
- 7.3 Gaussian Process and Initialisation 52
- 7.4 Bayesian Acquisition and Parameter Updates 53

8 Results and Discussion **55**
8.1 Results 55
8.2 Discussion 58
8.3 Further Work 58

Bibliography **59**

List of Tables

7.1 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. 53

List of Figures

2.1 Some simple model schematics, with varying numbers of compartments: S (susceptable), E (exposed), I (infectious) and R (recovered). The force of infection λ_t is usually a function of I_t, depicted by the dashed red lines. μ and ν are natural birth and death rates respectively. γ 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 susceptable again or into lifelong immunity, whereas in the SI with demography model, it can be interpreted as the increase to the rate of death attributable to disease induced mortality. σ is the rate of progression from a state of latent infection to becoming infectious. 6

2.2 Solutions to the ordinary differential equations 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 12

3.1 The *P. vivax* (malaria) lifecycle. *P. falciparum* does not have a dormant liver hypnozoite stage. Created with BioRender.com. 13

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 susceptable and infected humans respectively, and S_{M} and I_{M} are the number of susceptable and infected mosquitos. The rate of human infection (λ_{H}) is dependant on I_{M}, and the rate of human infection (λ_{M}) is dependant on I_{H}. 15

3.3 Diagram for *P. vivax* model in a tropical setting described by White et al. 2016. S and I are the number of susceptable and infected humans and mosquitos (denoted by subscript M). $\lambda_{\mathrm{H}} = mabI_{\mathrm{M}}$ and $\lambda_{\mathrm{M}} = ac(I_0 + I_{\mathrm{L}})$ 16

3.4 Diagram for *P. vivax* model described by Champagne et al. 2022. $I_{\mathrm{total}} = I_0 + I_{\mathrm{L}}$. Since the mosquito dynamics have been removed, λ 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 by a Poisson distribution with $\text{Pois}(y_i(\boldsymbol{\theta}))$. 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, genetated as I_t/N from the $SEIR$ simulation in Figure 2.3. All parameters were considered known except for β. The least squares estimate (LSE) $\beta^{\text{LSE}} = 0.3516$ for β 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 $\hat{\beta} = 0.3493$ for β was found by assuming the prevalence (times 1000) was binomially distributed from 1000 samples with the probability of success being 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. 26

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 β, 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 distibution $\beta \sim \text{Gamma}(2,6)$, where $\text{E}(\beta) = 1/3$. Our proposal density was $N(\beta^*, 1/10)$, where β^* was the previous sample. 30

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 hyperparameters $\ell = 1$, and $\sigma_o^2 = 1$. One realisation is bolded. Samples for each kernel were generated from the same seed. 38

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

6.1 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 $\sigma_k^2 = 1.1$ 43

6.2 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 ε_i is i.i.d. $\mathcal{N}(0, \sigma_o^2)$ 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 $\sigma_k^2 = 1.1$ 44

7.1 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 1000, with 10 initial infections (both blood and liver stage). . . . 51

8.1 Four sample discrepancies $\ln \mathcal{D}(\boldsymbol{\theta})$ taken along 21 values of each parameter to be predicted between the lower and upper parameter bound. The Gaussian process predicts the mean of $\ln \mathcal{D}(\boldsymbol{\theta})$ after initialisation, 50 iterations of Bayesian acquisition, and 500 iterations of Bayesian acquisition. All parameters not in the slice are fixed at the true parameters that generated the synthetic observed data. The black lines are $\mathrm{E}(d_{\mathcal{GP}}(\boldsymbol{\theta}))$ and the red lines are sample realisations from $(d_{\mathcal{GP}}(\boldsymbol{\theta}))$ 56

8.2 Final synthetic likelihoods $\hat{L}(\boldsymbol{\theta})$ after 500 iterations of Bayesian acquisition. Each likelihood shows how that likelihood changes across the parameter space. 57

Chapter 1

Introduction

All the stuff that's really hard about calibrating malaria model parameters. Recent model don't even calibrate using their models.

Why are we looking at mathematical models for vivax malaria and why do we need to do more complicated parameter inference. Outline that the limitation is that it takes a long time to run simulations from complicated malaria models and we still need to fit to data.

The aim of this thesis was to investigate the use of Gaussian Processes to

Part I

Literature Review

Chapter 2

Epidemiological Modelling

In order to study the behaviour and characteristics of disease spread and eradication, compartmental epidemiological models have been developed. They seek to simplify the dynamics of a disease down to a mathematically representable form. Inference on these models allow for an understanding of how the modelled disease spreads, and allows an assessment of how effective differing disease interventions (such as treatments or vaccinations) may be without the need for 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 be only be in one of a finite number of states (which are called compartments). These compartments usually correspond to a state of disease. Some simple common compartments include:

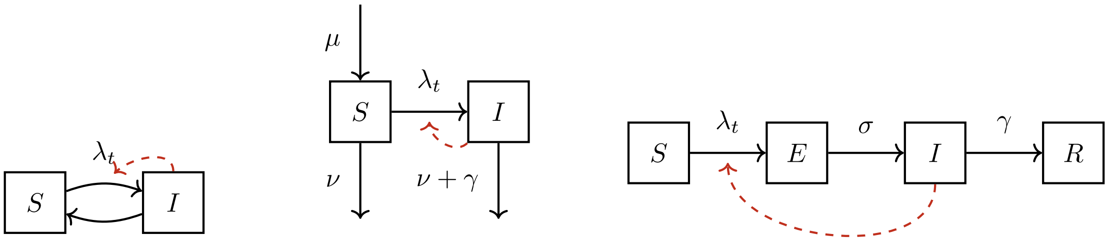
- S - Susceptable: at risk of contracting the disease
- E - Exposed: contracted the disease but not yet transmitting it
- I - Infectious (also called Infected): at risk of transmitting the disease
- R - Recovered: neither at risk of contracting or transmitting the disease.

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 (eg. S_t is the number of susceptibles at time t). Models are routinely described by the compartments they contain. For example, an SIS model, is a model with the susceptable and infectious compartments. Recovering from the infection leaves you susceptible to reinfection (for example most sexually transmitted diseases (Keeling and Rohani 2008, p. 56)), and is graphically depicted in Figure 2.1a.

Furthermore we can also include demography into a compartmental model. Diseases that infect the individual until the time of death such as bovine spongiform encephalopathy (BSE commonly known as mad cow disease) may be modelled using an SI model with demography (birth and death rates), as depicted in Figure 2.1b (Hagenaars, Donnelly, and Ferguson 2006).

Childhood diseases such as varicella (chickenpox) which give lifetime immunity after infection can be modelled using an $SEIR$ model (see Figure 2.1c), particularly when modelling a local outbreak setting (for example Zha et al. 2020 used the $SEIR$ model to model a school outbreak of varicella). Not including demograpy is usually appropriate when disease induced mortality is low.

The number (and names) of compartments can be extended and configured as needed, and compartments could be added for vaccinated individuals, quarantined individuals and so on. By



(a) SIS model

(b) SI with demography model

(c) $SEIR$ model

Figure 2.1: Some simple model schematics, with varying numbers of compartments: S (susceptable), E (exposed), I (infectious) and R (recovered). The force of infection λ_t is usually a function of I_t, depicted by the dashed red lines. μ and ν are natural birth and death rates respectively. γ 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 susceptable again or into lifelong immunity, whereas in the SI with demography model, it can be interpreted as the increase to the rate of death attributable to disease induced mortality. σ is the rate of progression from a state of latent infection to becoming infectious.

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 ODE models

Diseases are often simulated as deterministic ordinary differential equations. For the examples below we assume that the force of infection λ_t is proportional to the number of people in I, such that $\lambda_t := \beta\frac{I_t}{N_t}$. β 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. This is sometimes called the effective contact rate. Therefore 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 in a way that they would pass on the infection. In different diseases β varies dramatically, as some diseases need prolonged exposure or sexual contact to transmit, whereas some are very highly transmittable, and so will have very low β. Implicitly there is also a (sometimes poor) assumption of complete uniformly random mixing of people. Note that for this thesis we assume that β is frequency dependent (people interact with the same number of people regardless of population size) as opposed to density dependent (people interact with a number of people proportional to population size, in which case $\lambda := \beta I_t$.).

The SIS model depicted in Figure 2.1a, the ordinary different equations (ODEs) governing the model could be

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\lambda S_t + \gamma I_t = -\beta\frac{I_t}{N}S_t + \gamma I \tag{2.1}$$

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} = \lambda S_t - \gamma I_t = \beta\frac{I_t}{N}S_t - \gamma I. \tag{2.2}$$

With a stated assumption that population size is closed, equation 2.1 fully describes the model.

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

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = \mu N_t - \lambda_t S_t - \nu I_t = \mu N_t - \beta \frac{I_t}{N_t} S_t - \nu I \tag{2.3}$$

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} = \lambda_t S_t - (\gamma + \nu) I_t = \beta \frac{I_t}{N_t} S_t - (\gamma + \nu) I. \tag{2.4}$$

Here is it important to note that N_t is not necessarily constant.

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

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\lambda_t S_t - \nu I \qquad = -\beta \frac{I_t}{N} S_t + \gamma I_t \tag{2.5}$$

$$\frac{\mathrm{d}E_t}{\mathrm{d}t} = \lambda_t S_t - \omega E_t \qquad = \beta \frac{I_t}{N} S_t - \omega I_t \tag{2.6}$$

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} = \omega E_t - \gamma I_t \tag{2.7}$$

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = \gamma I_t \tag{2.8}$$

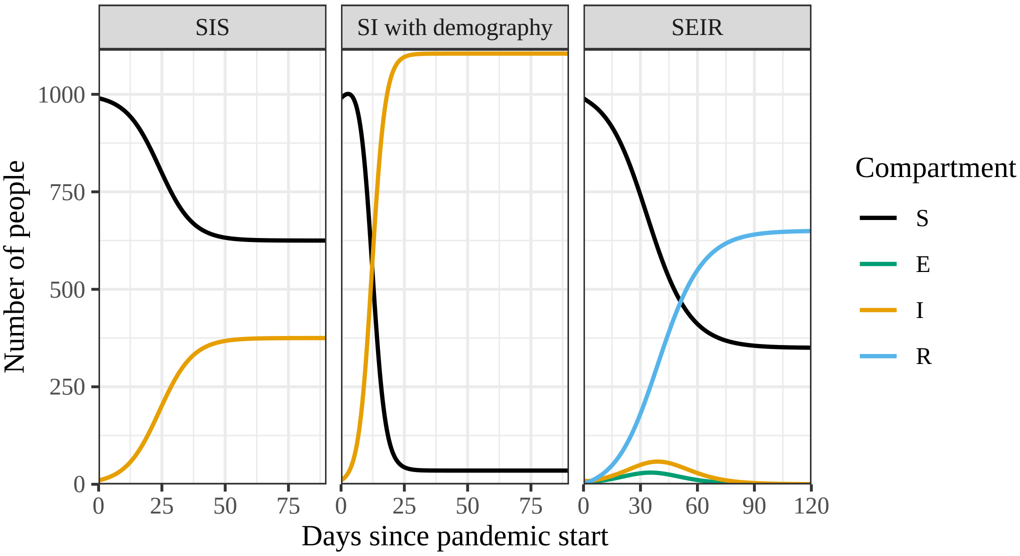


Figure 2.2: Solutions to the ordinary differential equations describing the models depicted in Figure 2.1. The initial infectious population was $I_0 = 10$, with $S_0 = 990$. In the $SEIR$ model, $E_0 = R_0 = 0$. For all models $\beta = 0.4$. For the SIS and SI model with demography $\gamma = 1/4$. For 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 initial for each compartment the ordinary differential equations have a deterministic output, such as in 2.2.

2.2 Stochastic models

Motivating the form of the stochastic model

To establish a relationship between the deterministic and stochastic disease models, we first need to establish 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* λ *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. $\mathcal{N}(t_2) - \mathcal{N}(t_1) \sim \text{Pois}(\lambda(t_2 - t_1)), 0 \leq t_1 < t_2$.

Deterministic ODE models are appropriate to study a kind of aggregate disease spread behaviour, and well approximate real world behaviour when the numbers in each compartment are large, however at the start of an epidemic, when the number of infected individuals is small the behaviour of the epidemic may vary significantly. It is possible that if the average number of people that an infectious individual infects near the beginning of the epidemic (formally referred to as R_0) is close to 1, then the disease may die out or become stable. Under the deterministic SIS model described by equations 2.1 and 2.2, consider the model at time t^* the instantaneous rate at which S is decreasing is $\beta\frac{I_{t^*}}{N}S_{t^*}$. In other words, one individual leaves the S compartment every $\beta\frac{I_{t^*}}{N}S_{t^*}$ units of time. We can 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^*.

$$\frac{\mathrm{dE}(\mathcal{N}_1(t^*))}{\mathrm{d}t^*} = \lim_{\delta\to 0}\frac{\mathrm{E}\left(\mathcal{N}(t^*+\delta) - \mathcal{N}(t^*)\right)}{\delta} = \frac{\beta\frac{I_{t^*}}{N}S_{t^*}(t^*+\delta-t^*)\beta\frac{I_{t^*}}{N}S_{t^*}}{\delta} = \beta\frac{I_{t^*}}{N}S_{t^*}.$$

Under the same deterministic formulation of the model, the instantaneous rate into S at t^* is γI_{t^*}. Therefore as above we can construct a Poisson point process $\{\mathcal{N}_2(t-t^*)\}_{t\geq t^*}$ with rate γI_{t^*} describing the number of recoveries from I to S, with $\frac{\mathrm{dE}(\mathcal{N}_2(t^*))}{\mathrm{d}t^*} = \gamma I_{t^*}$. Combining the two processes, we can see that the rate of change in the average number of people in S is

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

Therefore we can create a stochastic model where the local average in each compartment matches the behaviour of the ODE model at the same state. We do this first by formulating the model as a random vector $\{\mathbf{C}_t\}_{t\geq 0} = \{C_1(t), C_2(t), \dots, C_n(t)\}_{t\geq 0}$ where $C_i : \mathbb{R} \to \mathbb{N} \cup \{0\}$, 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 in a model with S and I compartments, $\{\mathbf{C}_t\}_{t\geq 0} := \{S_t, I_t\}_{t\geq 0}$. $\{\mathbf{C}_t\}_{t\geq 0}$ is a continuous time Markov chain (see Definition **??**) with transition kernel corresponding to the rates of the model. For example, in the SI model with demography in Figure 2.1b the transition rates are:

- $\{s, i\}$ to $\{s+1, i\}$ has rate $\mu(s+i)$
- $\{s, i\}$ to $\{s-1, i\}$ has rate νs

- $\{s, i\}$ to $\{s-1, i+1\}$ has rate $\beta\frac{i}{i+s}s$
- $\{s, i\}$ to $\{s, i-1\}$ has rate $(\nu+\gamma)i$.

We can interpret each transition as a seperate events, each behaving as independent Poisson point processes until the time of the first transition. Therefore at time t^* we have the Poisson point processes:

- $\{\mathcal{E}_1(t)\}_{t\geq 0}$: the number of births into S after time t^* with intensity μN_{t^*}
- $\{\mathcal{E}_2(t)\}_{t\geq 0}$: the number of deaths in S after time t^* with intensity νS_{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 $(\nu+\gamma)I_{t^*}$.

Theorem 2.2 (Sums of Independent Poisson Point Processes)**.** *Given independent Poisson point processes* $\{\mathcal{N}_1(t)\}_{t\geq 0}, \{\mathcal{N}_2(t)\}_{t\geq 0}, \ldots, \{\mathcal{N}_n(t)\}_{t\geq 0}$, *with intensities* $\lambda_1, \lambda_2, \ldots, \lambda_n$,

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

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

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

1. $\mathcal{N}(0) := \mathcal{N}_1(0) + \mathcal{N}_2(0) + \cdots + \mathcal{N}_n(0) = 0$ since $\mathcal{N}_i(0) = 0$ by definition of a Poisson point process.

2. We show that $\mathcal{N}(t_1)-\mathcal{N}(t_0), \mathcal{N}(t_2)-\mathcal{N}(t_1), \ldots, \mathcal{N}(t_n)-\mathcal{N}(t_{n-1})$ with $0 \leq t_0 < t_1 < \cdots < t_n$ are independent.

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

   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 $\mathcal{N}(t_1) - \mathcal{N}(t_0), \mathcal{N}(t_2) - \mathcal{N}(t_1), \ldots, \mathcal{N}(t_n) - \mathcal{N}(t_{n-1})$ with $0 \leq t_0 < t_1 < \cdots < t_n$ are independent.

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

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

   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

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

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 2.3 (Time to First Event in Poisson Point Process)**.** *Given a Poisson point process* $\{\mathcal{N}(t)\}_{t\geq 0}$ *with intensity* λ, *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.

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

□

By Theorem 2.2 and Theorem 2.3,

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

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 to the next event is random variable distributed

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

Theorem 2.4 (Probability of *i*th Poisson Process Generating the Next Event)**.** *Consider independent Poisson point processes*

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

having intensities $\lambda_1, \lambda_2, \ldots, \lambda_n$. *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 2.3, $\tau_i \sim \mathrm{Exp}(\lambda_i)$. Therefore

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

□

2.3 Doob-Gillespie Algorithm

All of this leads naturally to a common method of simulating the stochastic model. The Doob-Gillespie algorithm (often just called the Gillespie algorithm) is an algorithm that simulates a stochastic realisation of a model given a set of starting conditions.

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 $\lambda_i$ for all possible events $\mathcal{E}_i$
    Calculate total intensity $\lambda = \sum_i \lambda_i$
    Generate $\Delta t \sim \mathrm{Exp}(\lambda)$
    Choose event $\mathrm{E}_i$ with probability $\frac{\lambda_i}{\lambda}$
    Update time $t \leftarrow t + \Delta t$
    Update state of $\mathbf{C}(t + \delta t) \leftarrow \mathbf{C}(t) +$ change in state due to event $\mathcal{E}_i$
end while
```

2.4 τ-leaping

τ-leaping exploits the local Poisson point process like behaviour of epidemiological models. Consider the SIS model, when $S_t = I_t = 10000$. Events happen at a very high rate, meaning the Δt 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 τ as a Poisson point process with the total intensity $\lambda = \sum_i \lambda_i$ at time t, with the probability of any one event having the same probability as above of $\frac{\lambda_i}{\lambda}$. Therefore we have the following algorithm.

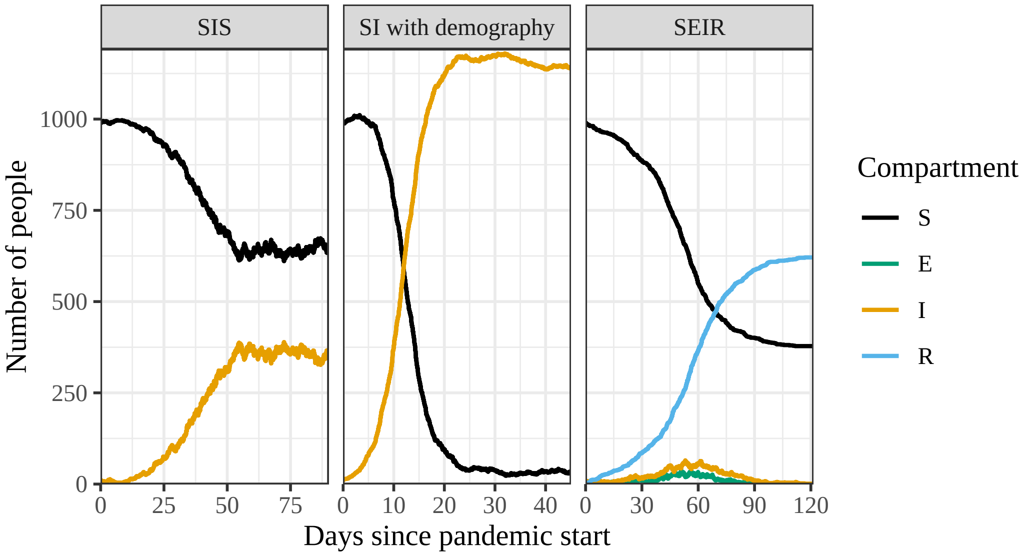


Figure 2.3: Exact stochastic simulations of the 3 different models using Algorithm 1. The parameters used were identical to those in Figure 2.2

Algorithm 2 τ-Leaping Algorithm

```
Initialise time $t \leftarrow 0$ and initial state of the model $\mathbf{C}(0) := \{C_1(0), C_2(0), \ldots, C_n(0)\}$
while termination condition not met do
    Calculate intensities $\lambda_i$ for all possible events $\mathcal{E}_i$
    Calculate total intensity $\lambda = \sum_i \lambda_i$
    Choose a suitable time step $\tau$ (this can be deterministic or adaptive)
    Calculate Poisson random variable $X \sim \text{Poisson}(\lambda)$
    for $i$ in 1 to $X$ do
        Choose event E$_i$ with probability $\frac{\lambda_i}{\lambda}$
        Update state of $\mathbf{C}(t+\tau) \leftarrow \mathbf{C}(t) +$ change in state due to event $\mathcal{E}_i$
    end for
    Update time $t \leftarrow t + \tau$
end while
```

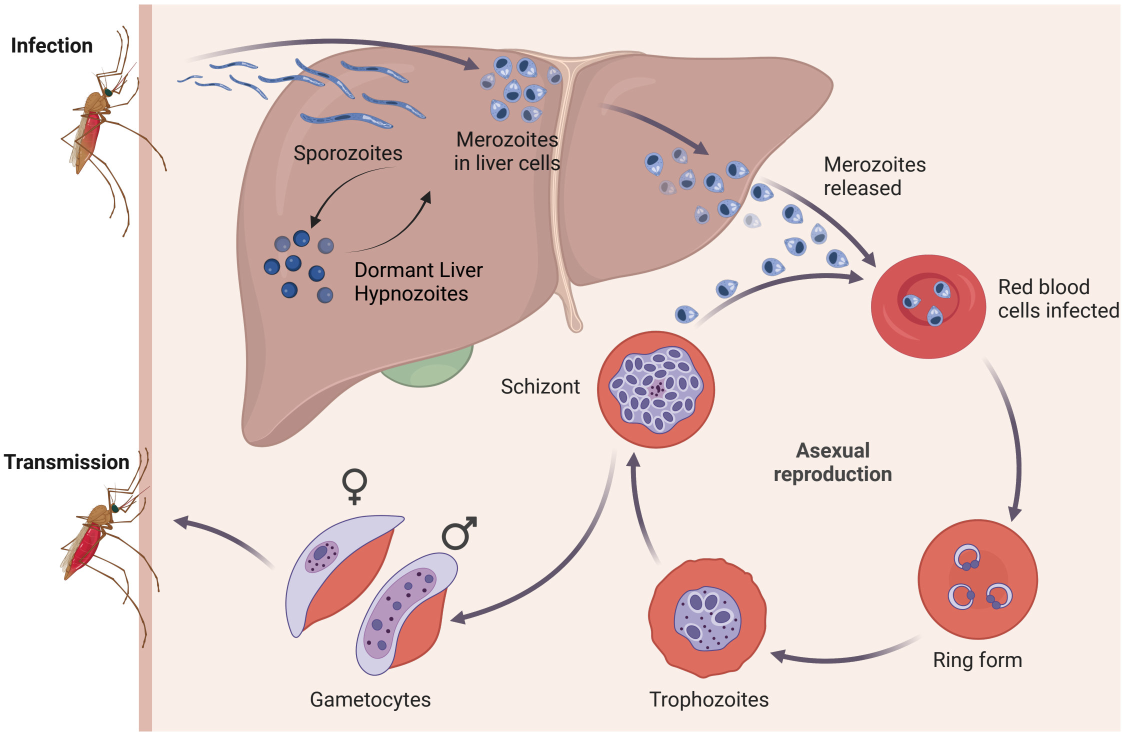
Chapter 3

Malaria and Malaria Models

3.1 Malaria

Malaria kills around 600,000 people each year, with over 75% of deaths occuring in children under 5 years old (World Health Organization 2022).

Plasmodium Life Cycle



Infection
Sporozoites
Merozoites in liver cells
Dormant Liver Hypnozoites
Merozoites released
Red blood cells infected
Schizont
Asexual reproduction
Transmission
♀
♂
Gametocytes
Trophozoites
Ring form

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

Malaria is a vector borne disease, needing both human (or other vertebrate) and mosquito hosts to complete it's lifecycle (Figure 3.1). Six species of the unicellular parasite 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). Sporozoites (a stage of the malaria parasite) enter the human blood stream via the skin after the female mosquito has a blood meal. From the blood stream they proceed to enter into the liver. Once a hepatocyte (liver cell) is invaded, the parasite will undergo asexual reproduction into up to 40,000 merozoites per hepatocyte, which are released into the blood stream . These merozoites then bind to, and invade erythrocytes (red blood cells), once again reproducing 16-32 fold in a process called schizogony. At this point, the erythrocyte membrane is ruptured, allowing for *Plasmodium* to invade new erythrocytes. Eventually, the merozoites undergo sexual differentiation, resulting in the sequestration and maturation of male and female gametocytes 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 when the mosquito next takes a blood feed (Cowman et al. 2016).

Illness, Treatment, and Immunity

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).

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).

Control and Eradication

Widespread use of DDTs in the mid 20th century led to significant successes in some countries towards the control and eradication of malaria. In the 1980s and 1990s, drug resistant malaria led to a doubling of malaria-attributable death. Currently, the control techniques include insecticide treated bed nets, and a mixture of antimalarial drugs (Cowman et al. 2016).

Plasmodium vivax

Unlike *P. falciparum*, *P. vivax* has hypnozoites, which are a dormant liver stage of the parasite. These can remain dormant for weeks and even months, leading to recurrent infections and illness, possibily 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 means diagnosis is more difficult, and it has an increased proportion of asymptomatic cases (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 old notion that *P. vivax* is benign has become untenable (Cowman et al. 2016).

Motivating Malaria Models

Levels of asymptomatic cases and latent 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, modelling malaria allows for modelling the effect of public health interventions such as mass treatment or testing, in order to determine an estimate of how effective the intervention may be, before large amounts of money are spent on trials.

3.2 Malaria Models

Ross-Macdonald

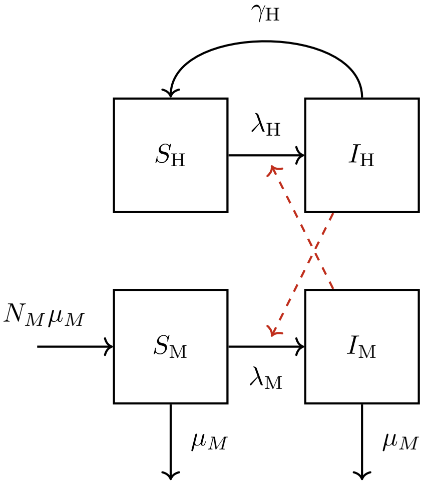


Figure 3.2: A simple Ross-Macdonald malaria model schematic, as described by Aron and May 1982. S_{H} and I_{H} are the number of susceptable and infected humans respectively, and S_{M} and I_{M} are the number of susceptable and infected mosquitos. The rate of human infection (λ_{H}) is dependant on I_{M}, and the rate of human infection (λ_{M}) is dependant on I_{H}.

Modelling malaria presents an additional challenge, as the disease is transmitted from mosquito to human and human to mosquito, rather than having direct human to human transmission. The most simple Ross-Macdonald model is depicted in figure 3.2. The ODEs for this model are

$$\begin{aligned}
\frac{\mathrm{d}S_{\mathrm{H}}}{\mathrm{d}t} &= \gamma_{\mathrm{H}} I_{\mathrm{H}} - bT_{\mathrm{HM}} I_{\mathrm{M}} \frac{S_{\mathrm{H}}}{N_{\mathrm{H}}} \\
\frac{\mathrm{d}I_{\mathrm{H}}}{\mathrm{d}t} &= bT_{\mathrm{HM}} I_{\mathrm{M}} \frac{S_{\mathrm{H}}}{N_{\mathrm{H}}} - \gamma I_{\mathrm{H}} \\
\frac{\mathrm{d}S_{\mathrm{M}}}{\mathrm{d}t} &= N_{\mathrm{M}}\mu_M + \gamma_{\mathrm{M}} I_{\mathrm{M}} - bT_{\mathrm{MH}} S_{\mathrm{M}} \frac{I_{\mathrm{H}}}{N_{\mathrm{H}}} - S_{\mathrm{M}}\mu_M \\
\frac{\mathrm{d}I_{\mathrm{M}}}{\mathrm{d}t} &= bT_{\mathrm{MH}} S_{\mathrm{M}} \frac{I_{\mathrm{H}}}{N_{\mathrm{H}}} - \gamma_{\mathrm{M}} I_{\mathrm{M}}
\end{aligned}$$

where b is the biting rate per mosquito, and T_{HM} is the probability of tranmission to a human given a bite by an infectious mosquito, with T_{MH} being vice-versa. Note that it is $\frac{I_{\mathrm{H}}}{N_{\mathrm{H}}}$ 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 don’t ‘recover’ from malaria due to their short lifespans, but the births and deaths are mathematically equivalent to assuming that the rate of ‘recovery’ amongst mosquitos is $\mu_{\mathrm{M}} I_{\mathrm{M}}$ per unit time, with no population dynamics.

A Ross-Macdonald style model simplifies 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

White Model

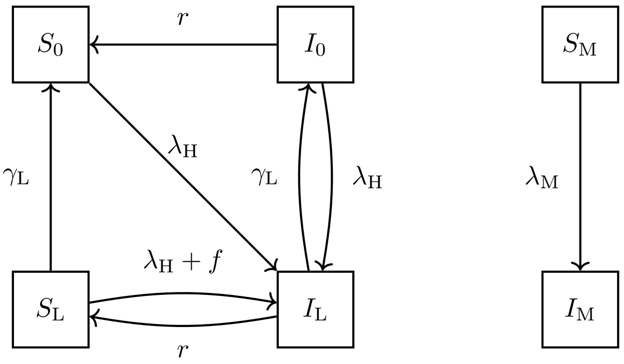


Figure 3.3: Diagram for *P. vivax* model in a tropical setting described by White et al. 2016. S and I are the number of susceptable and infected humans and mosquitos (denoted by subscript M). $\lambda_{\mathrm{H}} = mabI_{\mathrm{M}}$ and $\lambda_{\mathrm{M}} = ac(I_0 + I_{\mathrm{L}})$

The White model - described in White et al. 2016 (tropical model) and depicted in Figure 3.3 - is characterised by the following ordinary differential equations:

$$
\begin{aligned}
\frac{\mathrm{d}S_0}{\mathrm{d}t} &= -\lambda_{\mathrm{H}}S_0 + rI_0 + \gamma_{\mathrm{L}}S_{\mathrm{L}} \\
\frac{\mathrm{d}I_0}{\mathrm{d}t} &= -\lambda_{\mathrm{H}}I_0 - rI_0 + \gamma_{\mathrm{L}}I_{\mathrm{L}} \\
\frac{\mathrm{d}S_{\mathrm{L}}}{\mathrm{d}t} &= -\lambda_{\mathrm{H}}S_{\mathrm{L}} + rI_{\mathrm{L}} - fS_{\mathrm{L}} - \gamma_{\mathrm{L}}S_{\mathrm{L}} \\
\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}t} &= \lambda_{\mathrm{H}}(S_0 + I_0 + S_{\mathrm{L}}) - rI_{\mathrm{L}} + fS_{\mathrm{L}} - \gamma_{\mathrm{L}}I_{\mathrm{L}} \\
\frac{\mathrm{d}S_{\mathrm{M}}}{\mathrm{d}t} &= g - \lambda_{\mathrm{M}}(pS_{\mathrm{M}} - (1-p)I_{\mathrm{M}}) - gS_{\mathrm{M}} \\
\frac{\mathrm{d}I_{\mathrm{M}}}{\mathrm{d}t} &= \lambda_{\mathrm{M}}(pS_{\mathrm{M}} - (1-p)I_{\mathrm{M}}) - gI_{\mathrm{M}}. \qquad (I_0 + I_{\mathrm{L}} = \text{total number of bloodstage infections})
\end{aligned}
$$

It modifies the Ross-Macdonald models, to capture the differences in disease progression between *P. vivax* and *P. falciparum.* In particular, the White model includes the dormant liver stage that is unique to *P. vivax.*

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 λ_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. They can progress to either I_0 through the clearance of liver stage infection with rate γ_L, or to S_L through clearance of blood stage infection with rate r.

3. I_0 **(Infected Individuals - Blood-Stage Infection Only)** : Those in this compartment have a blood-stage infection with no latent hypnozoite infection in the liver. They are be reinfected into I_L with rate λ_H, relapse with rate f. Blood-stage infection is cleared (moving into compartment S_0) with rate r.

4. S_L **(Susceptable Individuals - Blood-Stage Infection Only)** : Those in this compartment have latent hypnozoite infection in the liver without blood-stage infection. They get novel infection through a mosquito bite into I_L with rate λ_H, or hypnozoite activation with rate f. This means that those in S_L move to compartment I_L with total rate $\lambda_\mathrm{H} + f$. Alternatively the hypnozoites are cleared from the liver (moving to compartment S_0) with rate γ_L.

5. S_M **(Susceptable Mosquitoes)** : Suspectable mosquitoes become infectious at rate $\lambda_\mathrm{M} p$. They die at rate $g + \lambda_\mathrm{M}(1-p)$. Since there is a constant mosquito population assumption, mosquitoes are born into this state at rate $g + \lambda_\mathrm{M}$.

6. I_M **(Infectious Mosquitoes)**: Infectious mosquitos die at rate $g + \lambda_\mathrm{M}(1-p)$.

$\lambda_\mathrm{H} := mabI_\mathrm{M}$ 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.

$\lambda_\mathrm{M} := ac(I_0 + I_\mathrm{L})$ 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 $V \sim \mathrm{Exp}(g)$, represent the lifespan of the mosquito. $\Pr(V > n) = 1 - F_V(n) = 1 - (1 - e^{-gn}) = e^{-gn}$.

$\lambda_\mathrm{M}(1-p)$ 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 mosquitoes. Presumably this approximates a model where mosquitoes are moved to an 'exposed' compartment for n time, after initial infection, however no justification is given in (White et al. 2016) 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

$$\frac{\mathrm{d}I_{\mathrm{M}}}{\mathrm{d}t} = \lambda^*_{\mathrm{M}} S_{\mathrm{M}} - gI_{\mathrm{M}} \quad \text{where } \lambda^*_{\mathrm{M}} := ac^*(I_0 + I_{\mathrm{L}}).$$

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 doesn't account for any importation of disease from an outside area, so if $S_0 = 1$, *P. vivax* is presumed permanently eradicated.

Champagne Model

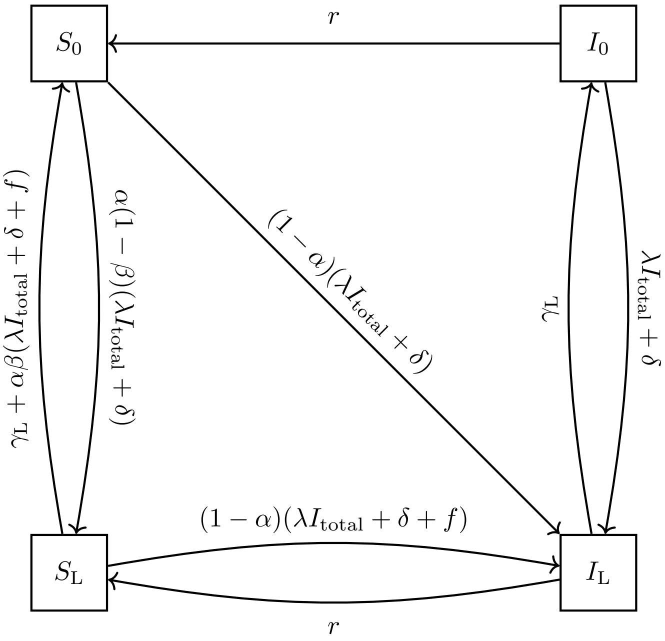

S_0
I_0
S_{L}
I_{L}
r
r
$\gamma_{\mathrm{L}} + \alpha\beta(\lambda I_{\mathrm{total}} + \delta + f)$
$\alpha(1-\beta)(\lambda I_{\mathrm{total}} + \delta)$
$(1-\alpha)(\lambda I_{\mathrm{total}} + \delta)$
γ_{L}
$\lambda I_{\mathrm{total}} + \delta$
$(1-\alpha)(\lambda I_{\mathrm{total}} + \delta + f)$

Figure 3.4: Diagram for *P. vivax* model described by Champagne et al. 2022. $I_{\mathrm{total}} = I_0 + I_{\mathrm{L}}$. Since the mosquito dynamics have been removed, λ 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 ordinary differential equations

$$\begin{aligned}
\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}t} =&(1-\alpha)(\lambda I_{\mathrm{total}}+\delta)(S_0+S_{\mathrm{L}})+(\lambda I_{\mathrm{total}}+\delta)I_0+(1-\alpha)fS_{\mathrm{L}}-\gamma_{\mathrm{L}}I_{\mathrm{L}}-rI_{\mathrm{L}}\\
\frac{\mathrm{d}I_0}{\mathrm{d}t} =&-(\lambda I_{\mathrm{total}}+\delta)I_0+\gamma_{\mathrm{L}}I_{\mathrm{L}}-rI_0\\
\frac{\mathrm{d}S_{\mathrm{L}}}{\mathrm{d}t} =&-(1-\alpha(1-\beta))(\lambda I_{\mathrm{total}}+\delta+f)S_{\mathrm{L}}+\alpha(1-\beta)(\lambda I_{\mathrm{total}}+\delta)S_0-\gamma_{\mathrm{L}}S_{\mathrm{L}}+rI_{\mathrm{L}}\\
\frac{\mathrm{d}S_0}{\mathrm{d}t} =&-(1-\alpha\beta)(\lambda I_{\mathrm{total}}+\delta)S_0+(\lambda I_{\mathrm{total}}+\delta)\alpha\beta S_{\mathrm{L}}+\alpha\beta fS_{\mathrm{L}}+\gamma_{\mathrm{L}}S_{\mathrm{L}}+rI_0
\end{aligned}$$

where $I_{\mathrm{total}} := I_0 + I_{\mathrm{L}}$.

The compartments have the same interpretation as in the White model, however the rates between compartments are significantly modified.

The new parameters are λ : the rate of infection, δ : importation rate, α : proportion of those infected who clear blood stage infections through immediate treatment, and β : 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 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 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 $\mathbf{y}(\boldsymbol{\theta}) = (y_1(\boldsymbol{\theta}), \dots, y_n(\boldsymbol{\theta}))$ 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 $\{(x_1, y_1(\boldsymbol{\theta})), \dots, (x_n, y_n(\boldsymbol{\theta}))\}$.

Least Squares Estimator

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

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

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

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

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* $\mathrm{E}(\varepsilon_i) = 0$. *We derive the least squares estimate of our parameters* $\boldsymbol{\theta} = (\theta_0, \theta_1)$ *by*

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

Since the expanded quadratic will have positive coefficients out the front of θ_0 *and* θ_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*

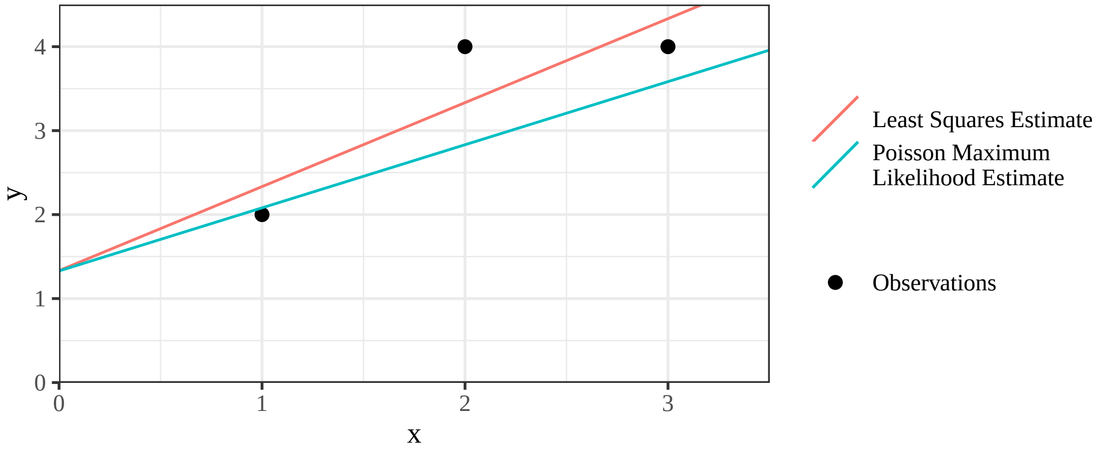


Figure 4.1: Two linear models of the form $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ fit given the set of observations $\{(1,2),(2,4),(3,4)\}$ using the method of least squares and maximum likelihood under the assumption that the data are independent realisations by a Poisson distribution with $\mathrm{Pois}(y_i(\boldsymbol{\theta}))$. 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 square method makes no explicit assumptions about the distribution of the noise ε. However if the distribution of ε 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* $\mathbf{y}^{\mathrm{obs}}$ *fixed, the* likelihood function *is*

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

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

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

The dependence on $\mathbf{y}^{\mathrm{obs}}$ is suppressed, but can be explicitly written as $\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}^{\mathrm{obs}})$. In the continuous (and mixture of discrete and continuous) case, we interpret $\Pr(\mathbf{y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} = \mathbf{y}^{\mathrm{obs}}|\boldsymbol{\theta})$ as the density $\Pr(\mathbf{y}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \in \mathrm{d}\mathbf{y}^{\mathrm{obs}}|\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ maximises the probability of the data. 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 ln 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^{obs} *were generated independently from* $y_i(\boldsymbol{\theta}) + \varepsilon_i \sim \mathrm{Pois}(y_i(\boldsymbol{\theta}))$*, where* $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ *as previously defined. Therefore the maximimum likelihood estimate of* $\boldsymbol{\theta}$ *is*

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

which we numerically solve to get $\hat{\theta}_0 \approx 1.329$ *and* $\hat{\theta}_1 \approx 0.751$*, as seen in Figure 4.1.*

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^{obs}s were has been generated with normal error.

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

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{\mathrm{LSE}}$$

Proof.

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

□

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 parameter to fit, the ODE model is fit to that data point. For example Champagne et al. 2022 fit one unknown model parameter to incidence 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 for using least square estimates. Another alternative 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) 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. Other data such as samples from the population to estimate prevalence (proportion of those infectious) could be distributed $\text{Binom}(n, \frac{I_t}{N})$.

We demonstrate estimation of one unknown parameter β from an $SEIR$ model, using prevalence data in Figure 4.2. Both β^{LSE} and $\hat{\beta}$ 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 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, $\boldsymbol{\theta}$ is considered to be fixed, with the observed data $\mathbf{y}^{\text{obs}}$ assumed to be generated from a distribution depending on $\boldsymbol{\theta}$. 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

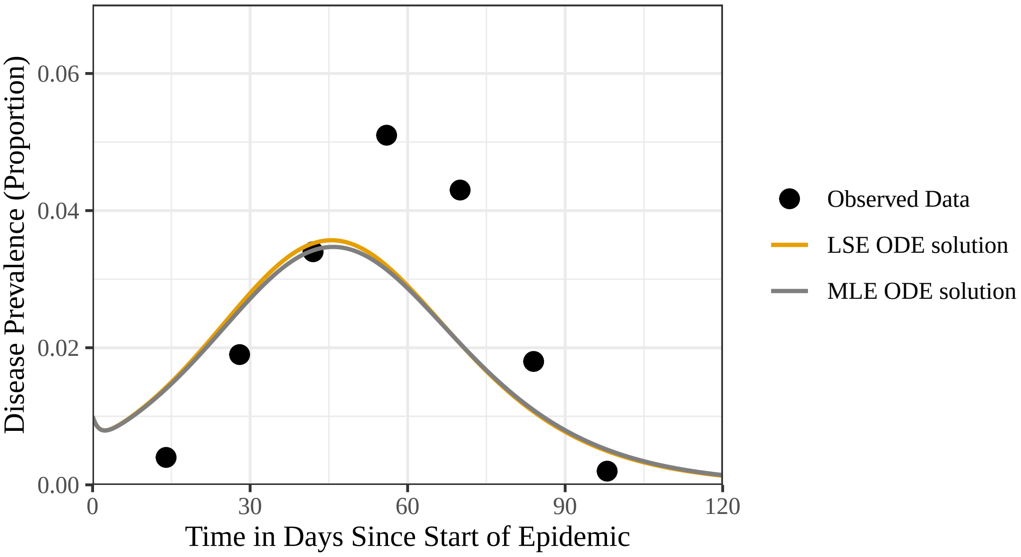


Figure 4.2: An $SEIR$ model fit to some observed prevalence data taken every two weeks over a 14 week period, genetated as I_t/N from the $SEIR$ simulation in Figure 2.3. All parameters were considered known except for β. The least squares estimate (LSE) $\beta^{\text{LSE}} = 0.3516$ for β was found 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 $\hat{\beta} = 0.3493$ for β was found by assuming the prevalence (times 1000) was binomially distributed from 1000 samples with the probability of success being equal to $\frac{I_t}{N}$.

$\boldsymbol{\theta}$ is also a random variable according to some pre-known prior distribution. 'Evidence' from the observed data then updates belief about $\boldsymbol{\theta}$, resulting in a posterior distribution of $\boldsymbol{\theta}$, described by Bayes' theorem, namely

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

Bayesian parameter estimation is still dependent on the likelihood function $\mathcal{L}(\boldsymbol{\theta}) := \Pr(\mathbf{y}^{\text{obs}}|\boldsymbol{\theta})$. With samples from the posterior distribution, we are able to run our models to capture uncertainty in predicting future scenarios. 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 the 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.

Rejection Sampling

If we have known ways to sample from the posterior parameter distribution, then we can use these. However if we have an equation for the probability distribution, but no way of sampling directly we can sample using rejection sampling. To sample X through rejection sampling, we need an explicit way of calculating $g(x)$ (where g is proportional to the density of X), a constant M and a distribution p such that $Mp(x) \geq g(x)$ with a sampling method available.

Under this methodology, the distribution function of X is

Algorithm 3 Rejection Sampler

```
Sample $X^* \sim p$
Sample $U \sim \mathrm{Unif}(0,1)$
if $U \leq \frac{g(X^*)}{Mp(X^*)}$ then
    return $X^*$ as a sample from the distribution of $X$
else
    Repeat
end if
```

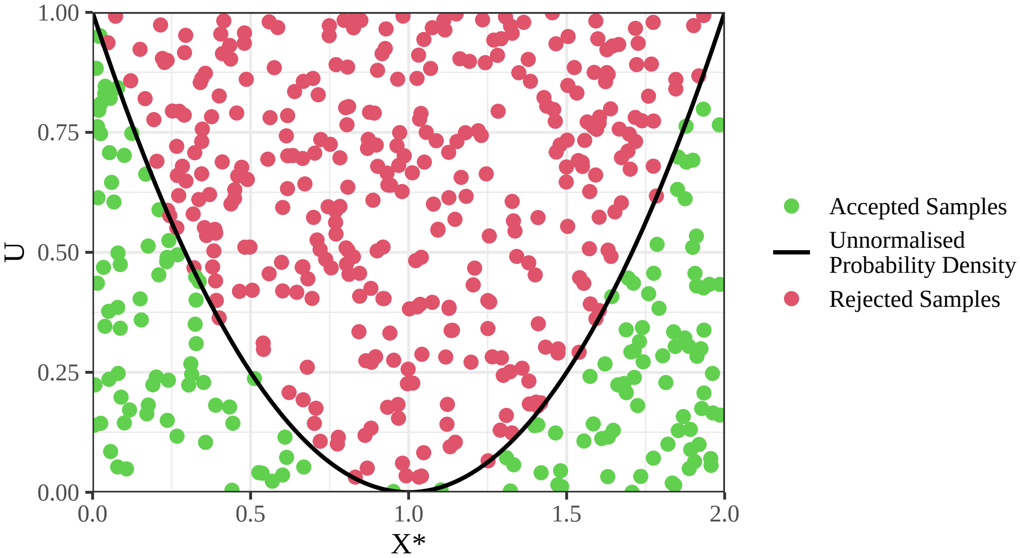


Figure 4.3: Samples of X from the unnormalised density $g(x) = (x-1)^2$ with $x \in (0,2)$ using the rejection sampler. $X^* \sim \mathrm{Unif}(0,2)$ and $M = 1$. Green dots are samples from from X. Of 500 samples of X^*, 157 were accepted as samples of X.

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

as required.

Example 4.7. *Let* $g(x) = (x-1)^2$ *be an unnormalised density function for* $x \in (0,2)$*.* $g(x) \leq 1$*, the density of a* $\mathrm{Unif}(0,2)$ *random variable. Therefore to generate samples from* g *we sample uniformly from* $X^* \sim \mathrm{Unif}(0,2)$*, and then accept the sample if a new* $U \sim \mathrm{Unif}(0,1)$ *is less than* $(X^* - 2)^2$*. This is demonstrated in 4.3*

Markov Chain Monte Carlo Methods

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

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

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

If $X_i \in \mathcal{X}$, then $\mathcal{X}$ is the state space of the Markov chain. For example the Markov chain constructed in Figure 4.4, has discrete state space $\mathcal{X} = \{1, 2\}$.

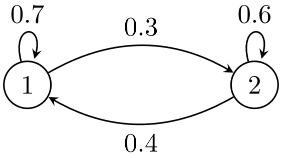


Figure 4.4: A simple time homogeneous Markov chain, with two states. It is characterised by the transition kernel $K(1,1) = \Pr(X_{i+1} = 1|X_i = 1) = 0.7$, $K(1,2) = \Pr(X_{i+1} = 2|X_i = 1) = 0.3$, $K(2,1) = \Pr(X_{i+1} = 1|X_i = 2) = 0.4$, and $K(2,2) = \Pr(X_{i+1} = 2|X_i = 2) = 0.6$. The stationary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

We will restrict our focus to 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*

$$\{X_i, X_{i+1}, \ldots, X_{i+n}\} \overset{d}{=} \{X_{i'}, X_{i'+1}, \ldots, X_{i'+n}\}$$

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 doesn't 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 π *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 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 which we do not discuss here. For a full discussion and proof see Robert and Casella 2010, chapter 6. □

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, then X' is accepted as the next state in the distribution with some probability α, chosen in such a way that if $X_i \sim g$, then $X_{i+1} \sim g$. Formally this is set out in Algorithm 4.

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 $\{x_0, x_1, \ldots, x_N\}$
```

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

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 $\alpha(x',x) = 1$.

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

□

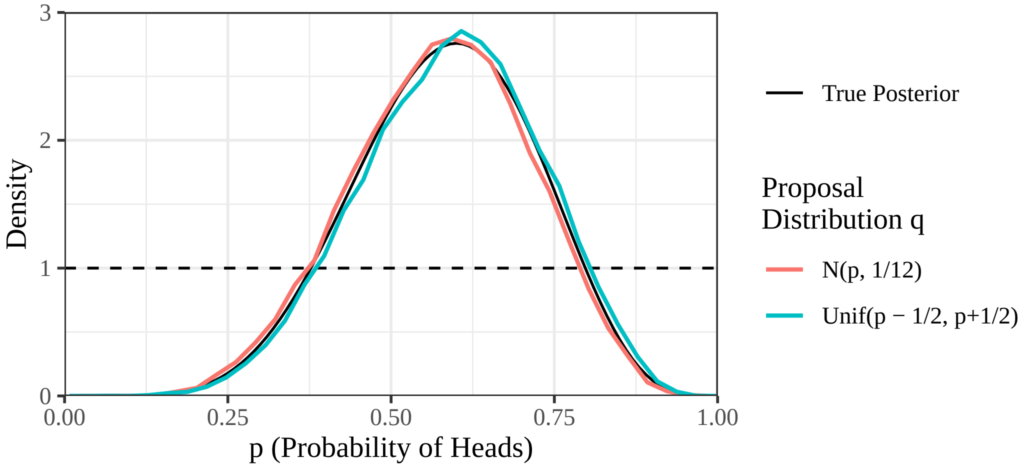


Figure 4.5: Samples from the posterior distribution of p using the Metropolis-Hastings algorithm. p was assumed to have a uniform prior between 0 and 1, with $y^{\text{obs}} = 6$, generated from $\text{Binom}(10, p)$, The choice of proposal distribution did not impact the final estimate of $\Pr(p|y^{\text{obs}})$.

Interestingly, the proof does not depend choice of proposal distribution, This can be seen in Example 4.14 and Figure 4.5.

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 $\alpha = \min\left(\frac{(P')^6(1-P')^4}{p_{i-1}^6(1-p_{i-1})^4}, 1\right)$        ▷ Assuming q symmetric
    Sample $U \sim \text{Uniform}(0,1)$
    if $U \le \alpha$ then
        $p_i \leftarrow X'$
    else
        $p_i \leftarrow x_{i-1}$
    end if
end for
```

12: ***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 Tth 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 are used (see Gelman et al. 2014, chapter 11).

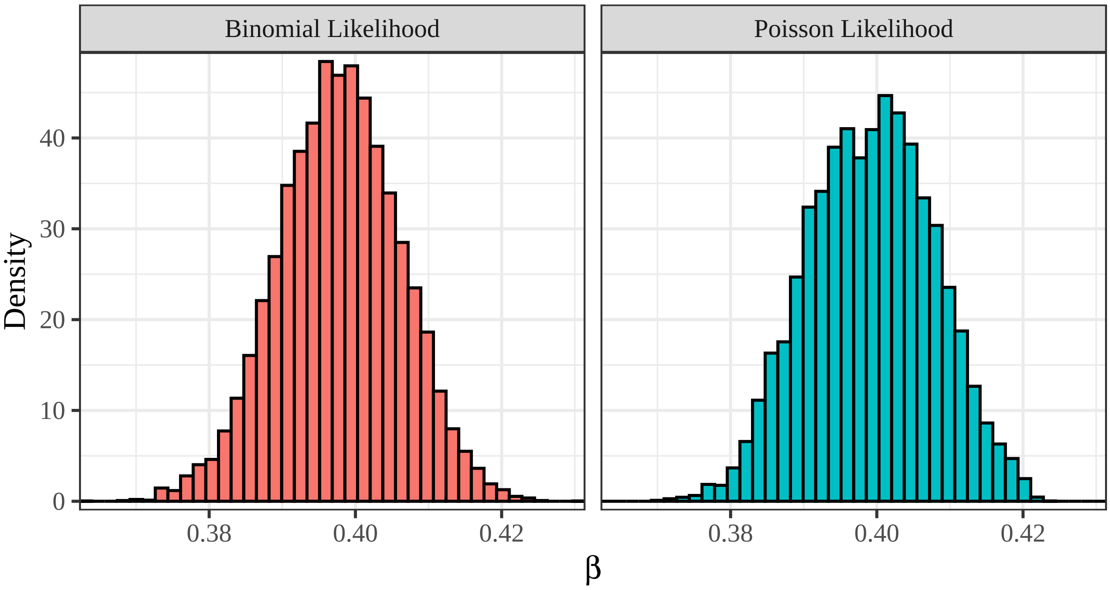


Figure 4.6: Given a daily incidence of $y^{\text{obs}} = 26$ at day 30 of an SIS epidemic, with unknown β, we use Metropolis-Hasting to sample from $\Pr(\beta|y^{\text{obs}})$. $\gamma = 1/4$ was assumed to be correct, and we compared the assumption $y^{\text{obs}} \sim \text{Binom}(\lfloor S_{30} \rfloor,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

For disease models, given a prior distribution for the parameter(s) $\boldsymbol{\theta}$, Metropolis-Hastings can be used to produce samples from $\Pr(\boldsymbol{\theta}|y^{\text{obs}}) \propto \mathcal{L}(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$, where $\mathcal{L}(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$ can be calculated to a proportionality constant but not directly sampled from. For example given an SIS model with unknown $\beta \sim \text{Gamma}(2, 6)$ and daily case counts y^{obs}, we can estimate sample from $\boldsymbol{\theta}|y^{\text{obs}}$ as in Figure 4.6.

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,

Algorithm 5 Gibbs Sampler

```
Initialise $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})$
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 $\boldsymbol{\theta}^{(i)}$
end for
return $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}\}$
```

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 satifies the detail 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

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

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

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

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*

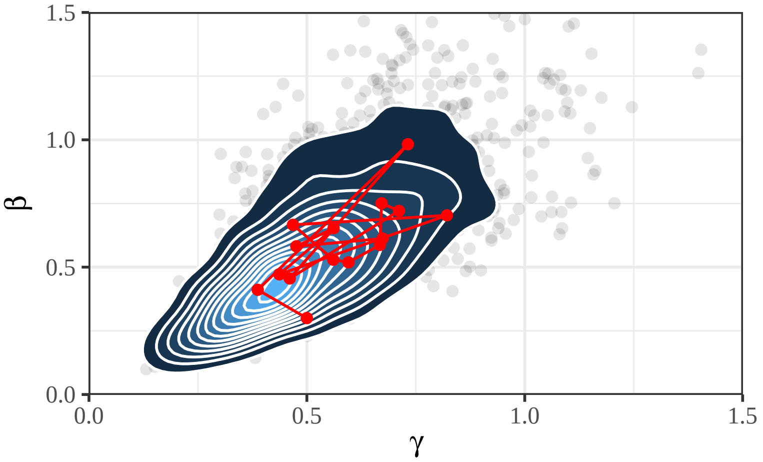


Figure 4.7: 2000 posterior samples from $\Pr(\beta, \gamma|\mathbf{y}^{\text{obs}})$, where $\beta|\gamma, \mathbf{y}^{\text{obs}} \sim \text{Gamma}(9, 4/\gamma + 4 + 8\gamma)$ and $\gamma|\beta, \mathbf{y}^{\text{obs}} \sim \text{InvGamma}(12, 12\beta)$. The samples were obtained using a Gibbs sampler.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

are generated from a Poisson distribution with mean $\frac{\beta}{\gamma}$*. Therefore the likelihood*

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

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 $\mathcal{L}(\boldsymbol{\theta}) := \Pr(\mathbf{y}^{\text{obs}}|\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$, and sample $\mathbf{y}(\boldsymbol{\theta})$ directly.

Algorithm 6 Naive Bayesian Sampler

```
Sample $\boldsymbol{\theta}^* \sim \Pr(\boldsymbol{\theta})$
Run model and compute $\mathbf{y}(\boldsymbol{\theta}^*)$
if $\mathbf{y}(\boldsymbol{\theta}^*) = \mathbf{y}^{\text{obs}}$ then
    return $\boldsymbol{\theta}^*$ as a sample from $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}}.)$
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, $\boldsymbol{\theta}^* = \mathbf{y}^{\text{obs}}$ 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 $\mathcal{L}(\boldsymbol{\theta}) := \Pr(\mathbf{y}(\boldsymbol{\theta}) \in \mathrm{d}\mathbf{y}^{\text{obs}}|\boldsymbol{\theta})$. Since

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

where $B_\epsilon^D(\mathbf{y}^{\text{obs}})$ is a ball of size ϵ 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 ϵ 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 $D(\mathbf{y}(\boldsymbol{\theta}), \mathbf{y}^{\text{obs}})$, are 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 $\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}^{\text{obs}}$

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

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 $||\mathbf{y}(\boldsymbol{\theta}) - \mathbf{y}^{\text{obs}}||_p$.

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 be random. . The full procedure is outlined in Algorithm 7.

Algorithm 7 Approximate Bayesian Compuation 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
```

Chapter 5

Gaussian Processes

5.1 Motivation and Definitions

If we knew the distribution of $\mathcal{D}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, then the need to sample from the model in Algorithm 7 becomes redundant, since we can sample directly from $\mathcal{D}(\boldsymbol{\theta})$. Moreover, the probability of drawing and accepting a sample from the becomes $\Pr(\boldsymbol{\theta})\Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon|\boldsymbol{\theta})$. Comparing this to Bayes' theorem, we can consider $\Pr(\mathcal{D}(\boldsymbol{\theta}) < \epsilon|\boldsymbol{\theta})$ an approximation of $\mathcal{L}(\boldsymbol{\theta})$ (up to some proportionality constant). We do not know the distribution of $\mathcal{D}(\boldsymbol{\theta})$, but we can construct an approximation. The approximation considered is achieved by modelling $\mathcal{D}(\boldsymbol{\theta})$ as a realisation of a Gaussian process.

The key insight that motivates our use of Gaussian processes is the assumption that 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.

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* $m : \mathcal{X} \to \mathbb{R}$ *and symmetric kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that for all finite sets* $\mathbf{x} := \{x_1, x_2, \ldots, x_n\} \subset \mathcal{J}$*, with* $f(\mathbf{x}) := [f(x_1), f(x_2), \ldots, f(x_n)]^T$

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

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

$$m(x_i) := \mathrm{E}\left[f(x_i)\right]$$

and

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

Although 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, the choice of covariance function $\mathrm{corr}(x, x') \to 1$ as $||x - x'|| \to 0$ induces continuity in x almost surely. Therefore they can be thought of as realisations of continuous functions. (Should I prove this?)

Some common examples of Gaussian processes include

1. Brownian motion on $\mathbb{R}$:
$$m \equiv 0, \quad \text{and} \quad k(s,t) = \min(s,t)$$

2. Ornstein Uhlenbeck process with parameters θ and σ:
$$m \equiv 0, \quad \text{and} \quad k(s,t) = \frac{\sigma_k^2}{2\theta}\left(e^{-\theta|t-s|} - e^{-\theta(t+s)}\right)$$

Properties such as the smoothness and undulation of the realised functions are also determined by the covariance kernel k, and associated hyperparameters. Before exploring different kernel options, we begin by defining a valid kernel function, and formalising 'smoothness.'

Definition 5.3 (Positive Semi-Definite Matrix)**.** *An* $n \times n$ *matrix* $\mathbf{A}$ *is* positive semi-definite *if* $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq 0$ *for all* $\mathbf{v} \in \mathbb{R}^n$.

Theorem 5.4 (Sufficient Condition for Positive Semi-Definite)**.** *A symmetric matrix* $\mathbf{A}$ *is positive semi-definite, if (and only if) it's eigenvalues are non-negative.*

Proof. □

Definition 5.5 (Positive Semi-Definite Kernel)**.** *A kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is* positive semi-definite *if the matrix*
$$\mathbf{K} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_n,x_1) & \dots & \dots & k(x_n,x_n) \end{bmatrix}$$
is positive semi-definite for any collection of $x_i \in \mathcal{X}$

Theorem 5.6. *All symmetric positive semi-definite matrices are covariance matrices for some set of random variables*

Proof. □

Definition 5.7 (Mean Square Continuous)**.** *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* mean square continuous *at* $\mathbf{x}$ *in the ith direction at if* $\mathrm{E}(|f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})|^2) \to 0$ *as* $|h| \to 0$, *where* $\mathbf{e}_i$ *is the unit vector with a 1 in the ith coordinate.*

Definition 5.8 (Mean Square Differentiable)**.** *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* mean square differentiable *at* $\mathbf{x}$ *in the ith direction with derivative* $\frac{\partial f(\mathbf{x})}{\partial x_i}$ *if*
$$\mathrm{E}\left[\left|\frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h} - \frac{\partial f(\mathbf{x})}{\partial x_i}\right|^2\right] \to 0$$
as $|h| \to 0$, *where* $\mathbf{e}_i$ *is the unit vector in the direction of the ith coordinate.*

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

Theorem 5.9. *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 $(B_{t+h} - B_t)^2 \sim |h|\chi_1^2 \to 0$ almost surely as $|h| \to 0$, and hence $\mathrm{E}[(B_{t+h} - B_t)^2 = 0]$. 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 $|h| \to 0$, as the variance approaches $+\infty$. ☐

Some common

Theorem 5.10 (Positive Semi-Definiteness of RBF)**.** *The radial basis function* $\sigma_k^2 \exp(-\frac{(x-x')^2}{2\gamma^2})$ *is a positive semi-definite kernel.*

Theorem 5.11 (Bochner's Theorem)**.** *Let* k *be a stationary kernel function such that* $k(x,x') = f(d)$*. A function* $k : \mathbb{R}^d \to \mathbb{C}$ *is the covariance function of a weakly stationary mean square continuous complex-valued random process of* $\mathbb{R}^d$ *if and only if it can be represented as*

$$k(\tau) = \int_{\mathbb{R}^d} \exp(2\pi i \boldsymbol{s} \cdot \tau)$$

(Rasmussen and Williams 2008, p. 82)

5.2 Families of Kernel Function

The two most common families of kernel functions are the squared exponential and Matérn families.

Matérn Family

The Matérn exponential kernel is of the form

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

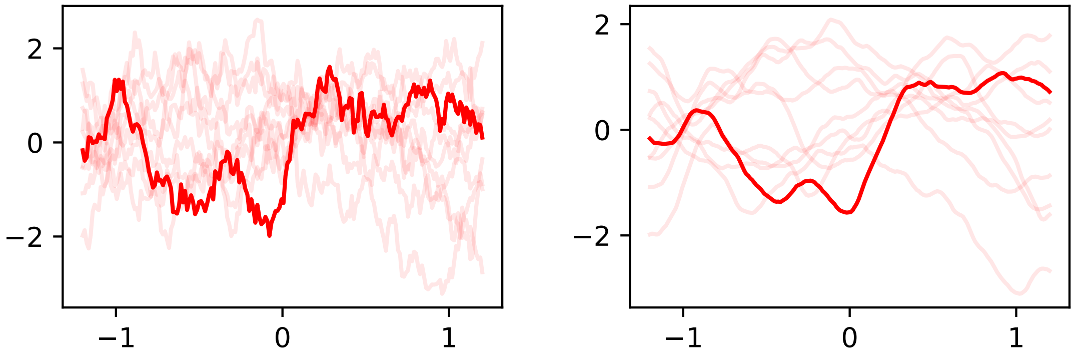
where K_ν is a modified Bessel function (defined in Abramowitz and Stegun 2013, p. 374). The general form is not very insightful, however for $\nu = 1/2, 3/2$ and $5/2$, (the most common values used) the kernel can be written as:

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

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

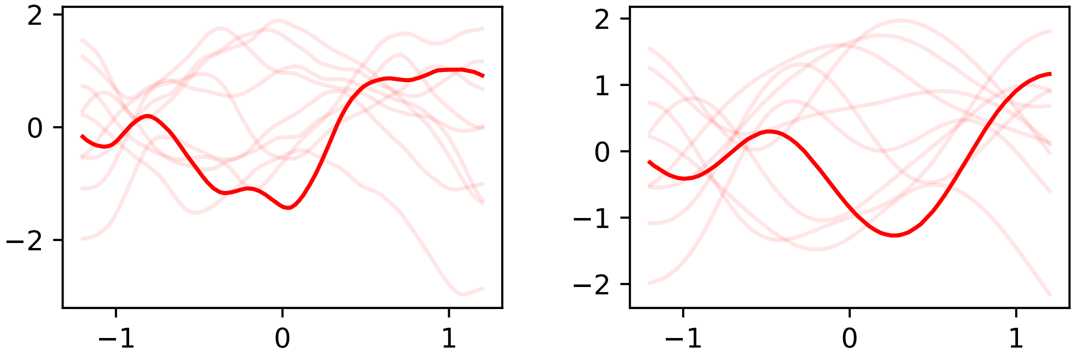
$$k_{5/2}(x,x') = \sigma_k^2 \left(1 + \frac{\sqrt{5}||x-x'||}{\ell} + \frac{5||x-x'||}{3\ell^2} \right) \exp\left(-\frac{||x-x'||^2}{2 * \ell^2} \right)$$

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. As $\nu \to \infty$, with appropriate rescaling, the limit of the Matérn kernel is the squared exponential kernel.Rasmussen and Williams 2008, p. 85 Proof in CHAPTER 4 SKOROKHOD STOCHASTIC I?



(a) Matérn 1/2 Kernel

(b) Matérn 3/2 Kernel



(c) Matérn 5/2 Kernel

(d) Squared Exponential Kernel

Figure 5.1: Ten sample realisations from 4 different kernels with hyperparameters $\ell = 1$, and $\sigma_o^2 = 1$. One realisation is bolded. Samples for each kernel were generated from the same seed.

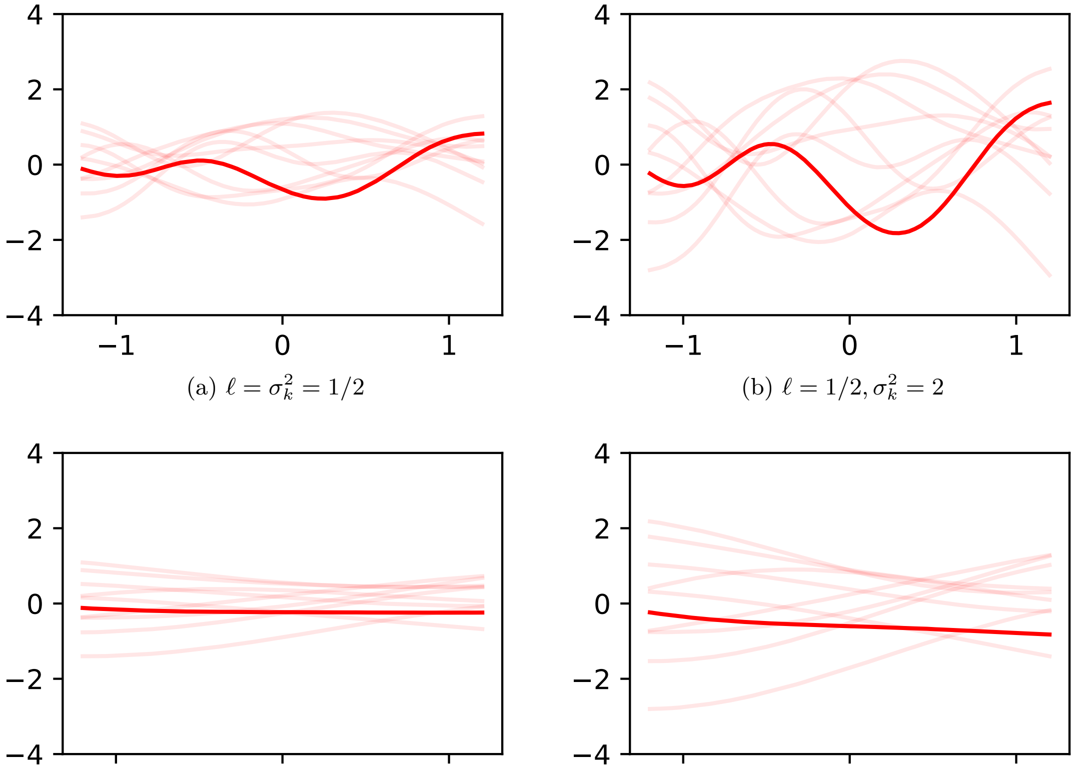
Squared Exponential Kernel

The squared exponential kernel has the form

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

As the limit of Matérn kernels, the squared exponential kernel is infinitely mean square differentiable. Despite this being the 'default' kernel in much of the literature, infinite differentiability is a very strong condition on functions which are very smooth, which can be seen in Figure 5.1d

Length and Amplitude Hyperparameters



(a) $\ell = \sigma_k^2 = 1/2$

(b) $\ell = 1/2, \sigma_k^2 = 2$

(c) $\ell = 2, \sigma_k^2 = 1/2$

(d) $\ell = \sigma_k^2 = 2$

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

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

Chapter 6

Gaussian Process Regression

Given the set of observations $f(\mathbf{x}_*)$ for the set of indices $\mathbf{x}_*$, it is often desirable to infer information about the function values at unobserved values. By choosing a Gaussian process with fixed kernel and hyperparameters, we can condition the process on the observed data, limiting the family of possible functions that we assume truly describe the model. Under the assumption that the function is a realisation of a Gaussian process predicting unseen function values reduces to elementary linear algebra. This is because a conditional multivariate normal distribution is still multivariate normal, and so the distribution of unobserved points will be multivariate normal.

Consider

$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \right)$$

Theorem 6.1 (Conditional Multivariate Normal Distribution is Multivariate Normal)**.** *With* $f(\mathbf{x})$ *and* $f(\mathbf{x}_*)$*, the conditional distribution is*

$$f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_* K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\ K - K_* K_{**}^{-1} K_*^T\right).$$

Proof. Since marginal distribution of the multivariate normal distribution, is also multivariate normal, $f(\mathbf{x}_*) \sim \mathcal{N}(m(\mathbf{x}_*), K)$. 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

$$
\begin{aligned}
p(f(\mathbf{x})|f(\mathbf{x}_*)) =& \frac{p(f(\mathbf{x}), f(\mathbf{x}_*))}{p(f(\mathbf{x}_*))}\\
\propto& \frac{\exp\left[-\frac{1}{2}\left(\begin{bmatrix} f(\mathbf{x})\\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x})\\ m(\mathbf{x}_*)\end{bmatrix}\right)^{\mathrm{T}} \begin{bmatrix} K & K_*\\ K_*^T & K\end{bmatrix}^{-1} \left(\begin{bmatrix} f(\mathbf{x})\\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x})\\ m(\mathbf{x}_*)\end{bmatrix}\right)\right]}{\exp\left[-\frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right]}\\
=& \exp\left[-\frac{1}{2}\left(\begin{bmatrix} f(\mathbf{x})\\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x})\\ m(\mathbf{x}_*)\end{bmatrix}\right)^{\mathrm{T}} \begin{bmatrix} K & K_*\\ K_*^T & K_{**}\end{bmatrix}^{-1} \left(\begin{bmatrix} f(\mathbf{x})\\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x})\\ m(\mathbf{x}_*)\end{bmatrix}\right)\right.\\
&\qquad \left. + \frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right]\\
=& \exp\left[-\frac{1}{2}\Big((f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}}\tilde{K}(f(\mathbf{x}) - m(\mathbf{x}))\right.\\
&\qquad + 2(f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}}\tilde{K}_*(f(\mathbf{x}_*) - m(\mathbf{x}_*))\\
&\qquad + (f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}}\tilde{K}_{**}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\Big)\\
&\qquad \left. + \frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right]\\
\propto& \exp\left[-\frac{1}{2}(f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}}\tilde{K}(f(\mathbf{x}) - m(\mathbf{x}))\right.\\
&\qquad \left. - (f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}}\tilde{K}_*(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right].\\
&\qquad\qquad\text{(by removing the terms independent of } f(\mathbf{x}))
\end{aligned}
$$

Since

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

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

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

and covariance matrix

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

by the alternative parametrisation 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,

$$
f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_*K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\ K - K_*K_{**}^{-1}K_*^T\right).
$$

□

After observing the function at multiple indices, we update the predictive distribution of any unobserved points, and generate new paths. The more points that the Gaussian process is conditioned on, the more narrow the sample paths, as seen in Figure 6.1

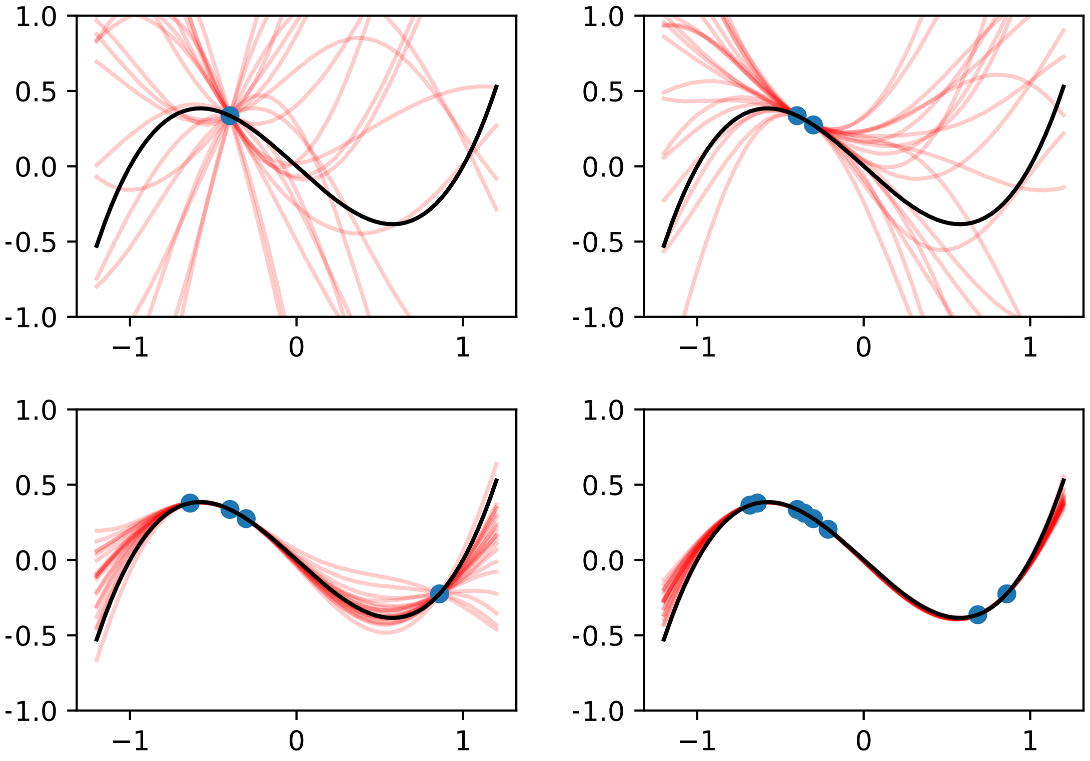


Figure 6.1: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 1, 2, 4, and 8 observations in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was zero mean and had a squared exponential kernel. The hyperparameters were fixed with $\ell = 2.7$ and $\sigma_k^2 = 1.1$

6.1 Observation Variance

For most functions, model outputs, or processes desirable for approximating through Gaussian process regression, multiple observations (through model runs or an real life measurements) of the same point will result in different observations. This is in contrast to exact realisations as in Figure 6.1. The simplest assumption is that the observations are of the form

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

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

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

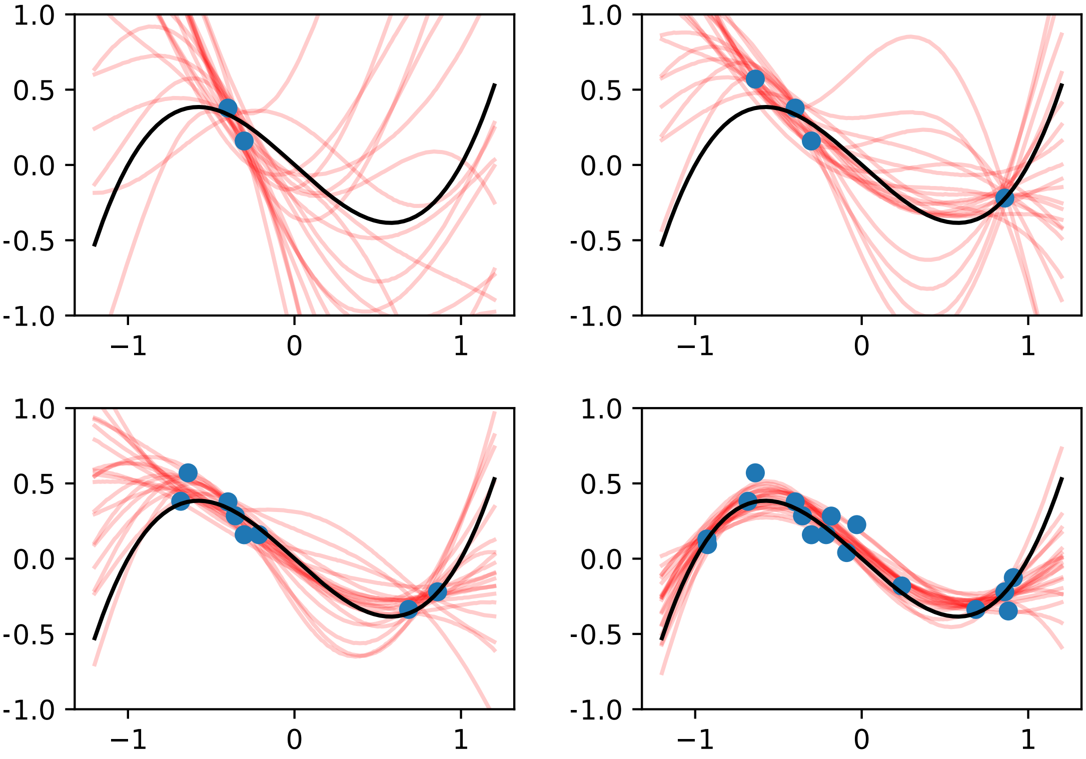


Figure 6.2: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 2, 4, 8, and 16 observations of $f(x_i) + \varepsilon_i$, where ε_i is i.i.d. $\mathcal{N}(0, \sigma_o^2)$ with $\sigma_o^2 = 0.01$ in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was 0 mean and had a squared exponential kernel. The hyperparameters were fixed with $\ell = 2.7$ and $\sigma_k^2 = 1.1$

Adding noise to the observations makes the predictive distributions means that a single observation gives less information about the underlying function, and hence the predictive distributions of unseen data are much less stable, as empirically seen in Figure 6.2.

6.2 Model Selection

Kernel Family

The appropriate choice of kernel will depend on the properties behaviour of the target function to approximate. In the case of estimating an extremely stochastic distribution (such as the price of a stock over time), it is unlikely to be smooth, so no mean square differentiability is required, and a Matérn 1/2 kernel would be appropriate. If it is known that our target function is extremely smooth, such as a finite sum of infinitely differentiable functions (such as polynomials, sin, cos etc.) then the choice of squared exponential kernel is the most appropriate kernel. Realistically, the smoothness of the function will not be known a priori, and hence some sort of compromise (such as Matérn 5/2) kernel allows for flexibility.

Many other kernels exist that induce varying behaviours, such as periodic kernels find periodic kernel, and non-stationary kernels (where the covariance is dependent on x and x', not just $|x-x'|$) have i defined stationary?.

Hyperparameters

The hyperparameters ℓ and σ_k^2 for a choice of kernel have to be are not known beforehand, unless the function is actually a realisation from a Gaussian process. Similarly, the observation variance σ_o^2 hyperparameter may not be a priori 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

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

which is equivalent to minimising

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

The covariance matrix generated by the choice of kernel K_{**} is explicitly written with its dependence on ℓ and σ_k^2. c is a constant.

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

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

where $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ is the distribution of the ith 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 6.1. Computationally efficient methods for calculating the predictive log probability 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).

Both methods can be extended to include estimating hyperparameters given a family of possible mean functions. For example Gutmann and Cor 2016 use maximimum likelihood estimates for the amplitude and turning points of the quadratic mean functions.

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.

Finally there is scope for a Bayesian approach to model selection. By setting priors on the hyperparameters to be estimated and using the likelihood as described in the maximum likelihood estimation approach, a posterior distribution can be easily contrived. A set of samples from this posterior could then be taken, or a less prinicipled maximum a posteriori probability estimate could then be taken for a point estimate of the hyperparameters. This approach has some obvious benefits, particularly when taking a posterior sample. Most obviously is that since multiple values for each hyperparameter are sampled, the set of functions after

6.3 Differing mean functions

Missing
figure

GP mean functions

6.4 Bayesian Acquisition Functions

Under the assumptions that making observations from the underlying function is costly, and we care about regions of the function with high (or low) values, new observations should be taken where there is high probability the function will be low. There also needs to be a trade-off between observing from areas with high predictive mean, and high predictive variance. These ideas are formalised by Bayesian acquisition functions $\mathcal{A}(x)$, with larger values corresponding to a higher 'desirability.' The target function is then sampled at the x which maximises this aquisition function. The new observation is then incorporated into the acquisition function.

Upper Confidence Bound

The upper confidence bound is one common way of exploring this trade off. The upper confidence bound

$$\mathcal{A}_{\mathrm{UCB}}(x) := \mathrm{E}[f(x)|f(\mathbf{x}_*)] + \eta_t\sqrt{\mathrm{Var}[f(x)|f(\mathbf{x}_*)]}$$

$$\mu(\boldsymbol{\theta}) - \eta_t\sqrt{\mathrm{v}(\boldsymbol{\theta})}$$

with $\eta_t := \sqrt{2\ln(\frac{t^{2d+2}\pi^2}{3\varepsilon})}$, and $\varepsilon \in (0,1)$ that can be chosen (with a lower epsilon resulting in more exploration), $\mu(\boldsymbol{\theta})$ and $\mathrm{v}(\boldsymbol{\theta})$ are the posterior mean and variance. $\varepsilon = 0.1$ was used.

Similarly we can also define

- BOLFI paper uses
$$\mu(\boldsymbol{\theta}) - \eta_t \sqrt{\mathrm{v}(\boldsymbol{\theta})}$$
    - $\eta_t := \sqrt{c + 2\ln(t^{d/2+2})}$, and c can be chosen
    - $\mu(\boldsymbol{\theta})$ and $\mathrm{v}(\boldsymbol{\theta})$ are the posterior mean and variance
- Could use expected information
$$(\mu_{\min} - \mu(\boldsymbol{\theta}))\Phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{\mathrm{v}(\boldsymbol{\theta})}}\right) + \sqrt{\mathrm{v}(\boldsymbol{\theta})}\phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{\mathrm{v}(\boldsymbol{\theta})}}\right)$$
    - $\mu_{\min} := \min_{\boldsymbol{\theta}} \mu(\boldsymbol{\theta})$
    - Φ, ϕ CDF and PDF of standard normal

exploration parameter proven by Srinivas et al. 2010

Show why Bayesian Acquisition fn not work for infinite space.

Part II

Calibrating Parameters for a *P. vivax* Model

Chapter 7

Methods

7.1 Creation of Synthetic Data

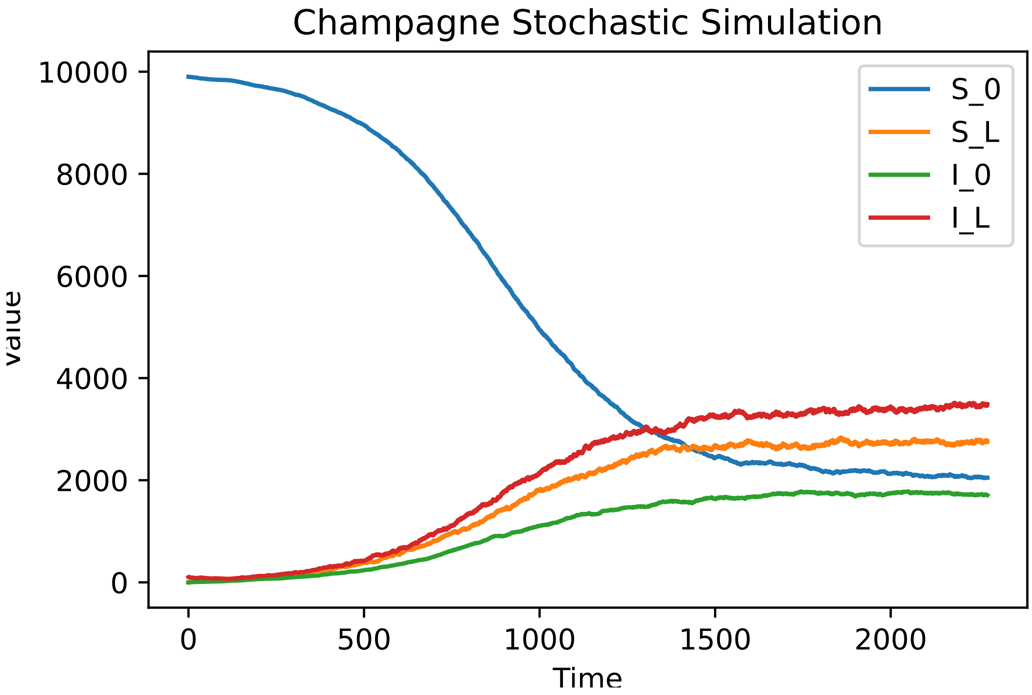


Figure 7.1: A Doob-Gillespie Simulation of the model described by Champagne et al. 2022 with $\alpha = 0.4$, $\beta = 0.4$, $\gamma_L = 1/223$, $\lambda = 0.04$, $f = 1/72$, $r = 1/60$, and $\delta = 0$. The population was 1000, with 10 initial infections (both blood and liver stage).

We investigated the model by Champagne et al. 2022 as described in 3.2. A malaria epidemic was simulated using the Doob-Gillespie algorithm (see Figure 7.1), using a population size of 1,000, and initial infected population of 10 (with both liver and blood stage infection). The parameters used closely followed those reported in Champagne et al. 2022:

- effective blood stage treatment proportion $\alpha = 0.4$,
- effective liver stage treatment proportion $\beta = 0.4$,

- rate of liver stage disease clearance $\gamma_L = 1/223$ days^{-1},
- importation rate $\delta = 0$ (assumed to be known),
- rate of infection $\lambda = 0.04$ days^{-1},
- rate of relapse $f = 1/72$ days^{-1},
- rate of blood stage disease clearance $r = 1/60$ days^{-1}.

From intialisation, the simulation was run for 15,000 events (with an event being anything that caused the size of any compartment to change such as an infection, recovery, relapse etc.), after which, the model was understood to have reached steady state behaviour.

The synthetic data (as summary statistics) measured from the steady state were:

1. p_{obs} : the number of currently infected individuals at the end of simulated epidemic (steady state prevalence).
2. m_{obs} : the number of cases in the first week of the epidemic (first month incidence)
3. w_{obs} : the number of cases in the last week of the epidemic (steady state weekly incidence).

New infections which instantly undergo radical cure don't change the size of each compartment. The number of these 'silent' incidences were calculated between events using a Poisson distribution with rate $\Delta t \times \alpha\beta\lambda(I_L + I_0)S_0/N$, where Δt is the time between events.

7.2 Model Simulations and Discrepancy Function

New epidemics were simulated as above, with 15,000 events and at least 30 days (to allow for calculation of incidence in the first month of the epidemic). For each model run steady state prevalence p, first month incidence m, and steady state weekly incidence w were calculated with the identical method to the synthetic data.

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

$$\mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r) := \sqrt{\left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2 + \left(\frac{m - m_{\text{obs}}}{m_{\text{obs}}}\right)^2 + \left(\frac{w - w_{\text{obs}}}{w_{\text{obs}}}\right)^2}.$$

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

7.3 Gaussian Process and Initialisation

We used a Gaussian process $d_{\mathcal{GP}}(\boldsymbol{\theta})$ as a surrogate model for $\mathrm{E}[\ln \mathcal{D}(\boldsymbol{\theta})]$. It was regressed on samples of $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$, where $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$ is the sample mean of 30 $\ln \mathcal{D}(\boldsymbol{\theta})$ samples. We used the kernel

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

where

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

that is, a Matérn kernel with $\nu = 5/2$ and automatic relevance determination - i.e. each parameter $\theta \in \boldsymbol{\theta}$ was scaled by ℓ_θ. This can be seen as giving each parameter its own length hyperparameter. The Gaussian process was constant mean $m_{\mathcal{GP}}$ with 'observation' variance σ_o^2. σ_o^2 is naturally intepretable as the sample mean variance.

Modelling the mean of the (log) discrepancy function is theoretically more sound than modelling the (log) discrepancy function itself, because we were not able to find theoretically sound reasons to assume normality, but the mean is asymptotically normal under reasonable conditions by the central limit theorem. For example, even under the simplest case that the discrepancy is L_2 norm of k differences x_i where $x_i \sim \mathcal{N}(0,1)$, $\sqrt{\sum_{i=1}^k x_i^2} \sim \chi(k)$, a Chi distribution with k degrees of freedom. do I need to cite a reference for this distribution (wikipedia)

Parameter	Upper Bound	Unit
Proportion of treatment clearing blood stage disease α	1	
Proportion of treatment clearing liver stage disease β	1	
Rate of liver stage disease clearance γ_L	1/30	1/days
Rate of infection λ	1/10	1/days
Rate of relapse f	1/14	1/days
Rate of blood stage disease clearance r	1/14	1/days

Table 7.1: 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.

All parameters to be calibrated were given conservative upper bounds after considering values reported in the literature, which informed where to define the Gaussian process. Parameter values outside this range were not considered.

Latin hypercube sampling was used to generate initialise 50 samples of the parameter space (scaled to be between zero and the upper bounds described in Table 7.1). For each set of parameters, $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$ was generated. The hyperparameters $m_{\mathcal{GP}}$, $\sigma_o^2, \sigma_a^2, \ell_\alpha, \ell_\beta, \ell_{\gamma_L}, \ell_\lambda, \ell_f$, and ℓ_r were selected by leave one out cross validation.

7.4 Bayesian Acquisition and Parameter Updates

For 500 iterations, the next $\boldsymbol{\theta}$ to sample $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$ from at was found by maximising the expected information acquisition function $\mathcal{A}_{\mathrm{EI}}(\boldsymbol{\theta})$. Let the current iteration be t, and $d_{\mathcal{GP}}^{(i)}(\boldsymbol{\theta})$ be the Gaussian process regressed on the simulated $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$s after i iterations.

Each index of $\boldsymbol{\theta}$ was initialised randomly at either the previous sample of $\boldsymbol{\theta}$ which minimised $\mathrm{E}(d_{\mathcal{GP}}^{(t-1)}(\boldsymbol{\theta}))$, or uniformly at random between it's lower and upper bounds. In each iteration, there was a $\min\left[1/5 + \exp(1 - t/4), 1\right]$ probability that one of the parameters (say θ^*) in $\boldsymbol{\theta}$ was chosen, and $\overline{\ln \mathcal{D}}(\boldsymbol{\theta})$ was sampled at $\boldsymbol{\theta}$ as well as at 11 evenly spaced values of θ^*, with the other parameters fixed. $1/5 + \exp(1 - t/4) > 1$ for small t, decaying to $1/5$ as t is large. This was to help initialise the Gaussian process model, as well as optimise the ℓ_θs. Every 50 iterations, the hyperparameters were reoptimised using leave one out cross validation. Finally, the synthetic likelihood was calculated using $\hat{\mathcal{L}}(\boldsymbol{\theta}) := \Pr(d_{\mathcal{N}}(\boldsymbol{\theta}) < \epsilon)$, where $d_{\mathcal{N}} \sim \mathcal{N}(\mathrm{E}[d_{\mathcal{GP}}^{(500)}(\boldsymbol{\theta})], 30^2\sigma_o^2)$

The Gaussian process and Gaussian process regression was implemented using TensorFlow in Python (Martín Abadi et al. 2015).

Chapter 8

Results and Discussion

8.1 Results

should these be presented in tables?

The synthetic data generated were

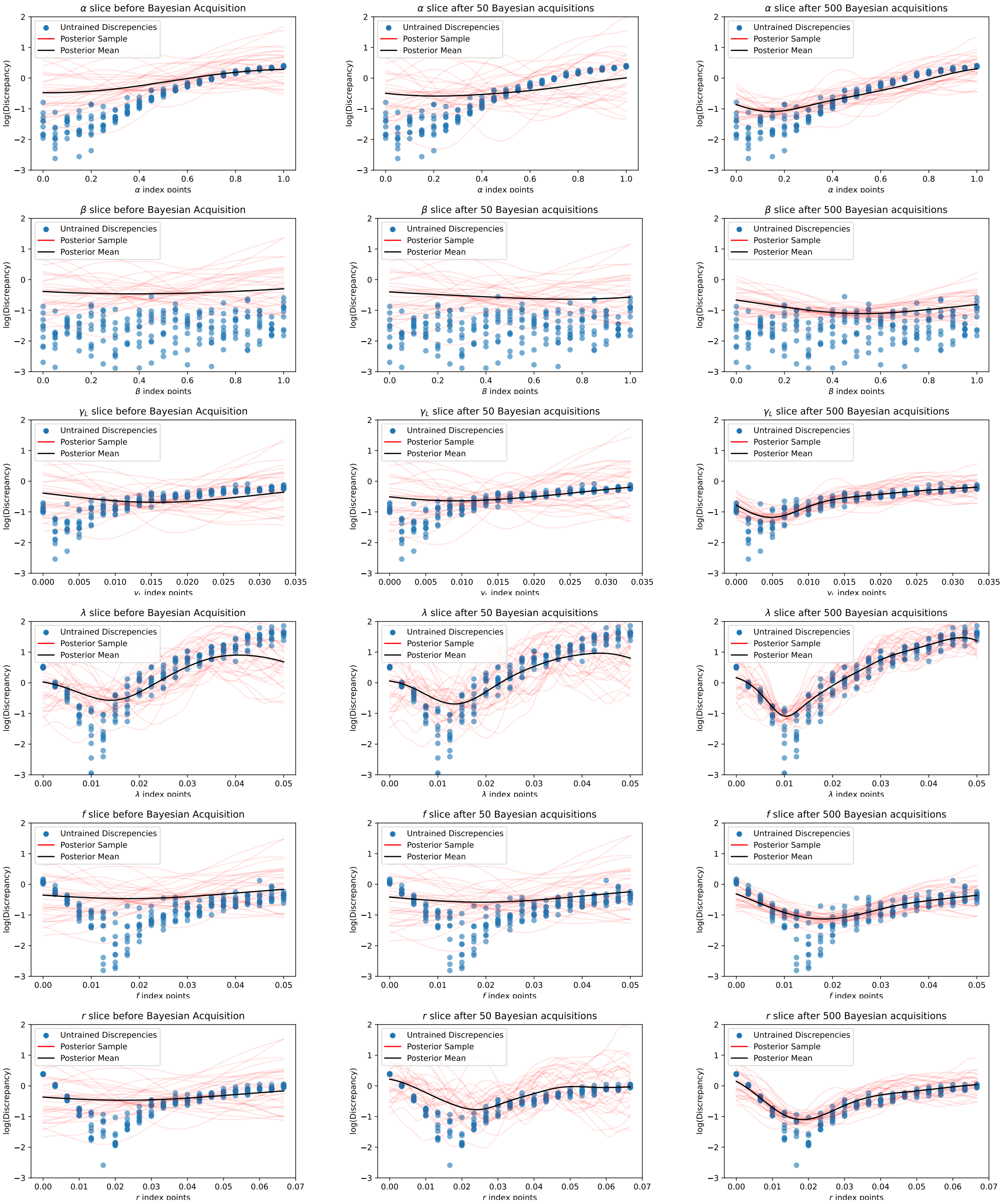
- $p_{\text{obs}} = 593$ final prevalence
- $m_{\text{obs}} = 13$
- $w_{\text{obs}} = 73$.

The final hyperparameter estimates were

- $\sigma_k^2 = 0.677$
- $\sigma_o^2 = 0.093$
- $\ell_\alpha = 0.232$
- $\ell_\beta = 0.25$
- $\ell_{\gamma_L} = 0.008$
- $\ell_\lambda = 0.015$
- $\ell_f = 0.018$
- $\ell_r = 0.013$

The minimum predicted mean of a sampled $\boldsymbol{\theta}$ was at $(0.492, 0.9, 0.029, 0.03, 0.064, 0.023)$, compared to the true parameters used to generate the data $(0.4, 0.4, 0.004, 0.04, 0.014, 0.017)$.

As the number of Bayesian acquisitions increased, the Gaussian process mean seems to converge to the true mean, as seen in Figure 8.1. The likelihood has maximum around the true values as seen in Figure 8.2. I will add in lines where the true values are to this figure. There seems to be a lack of good data



α slice before Bayesian Acquisition

α slice after 50 Bayesian acquisitions

α slice after 500 Bayesian acquisitions

β slice before Bayesian Acquisition

β slice after 50 Bayesian acquisitions

β slice after 500 Bayesian acquisitions

γ_L slice before Bayesian Acquisition

γ_L slice after 50 Bayesian acquisitions

γ_L slice after 500 Bayesian acquisitions

λ slice before Bayesian Acquisition

λ slice after 50 Bayesian acquisitions

λ slice after 500 Bayesian acquisitions

f slice before Bayesian Acquisition

f slice after 50 Bayesian acquisitions

f slice after 500 Bayesian acquisitions

r slice before Bayesian Acquisition

r slice after 50 Bayesian acquisitions

r slice after 500 Bayesian acquisitions

Untrained Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
2
1
0
−1
−2
−3
α index points
0.0 0.2 0.4 0.6 0.8 1.0
β index points
0.0 0.2 0.4 0.6 0.8 1.0
γ_L index points
0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035
λ index points
0.00 0.01 0.02 0.03 0.04 0.05
f index points
0.00 0.01 0.02 0.03 0.04 0.05
r index points
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07

Figure 8.1: Four sample discrepancies $\ln \mathcal{D}(\boldsymbol{\theta})$ taken along 21 values of each parameter to be predicted between the lower and upper parameter bound. The Gaussian process predicts the mean of $\ln \mathcal{D}(\boldsymbol{\theta})$ after initialisation, 50 iterations of Bayesian acquisition, and 500 iterations of Bayesian acquisition. All parameters not in the slice are fixed at the true parameters that generated the synthetic observed data. The black lines are $\mathrm{E}(d_{\mathcal{GP}}(\boldsymbol{\theta}))$ and the red lines are sample realisations from $(d_{\mathcal{GP}}(\boldsymbol{\theta}))$

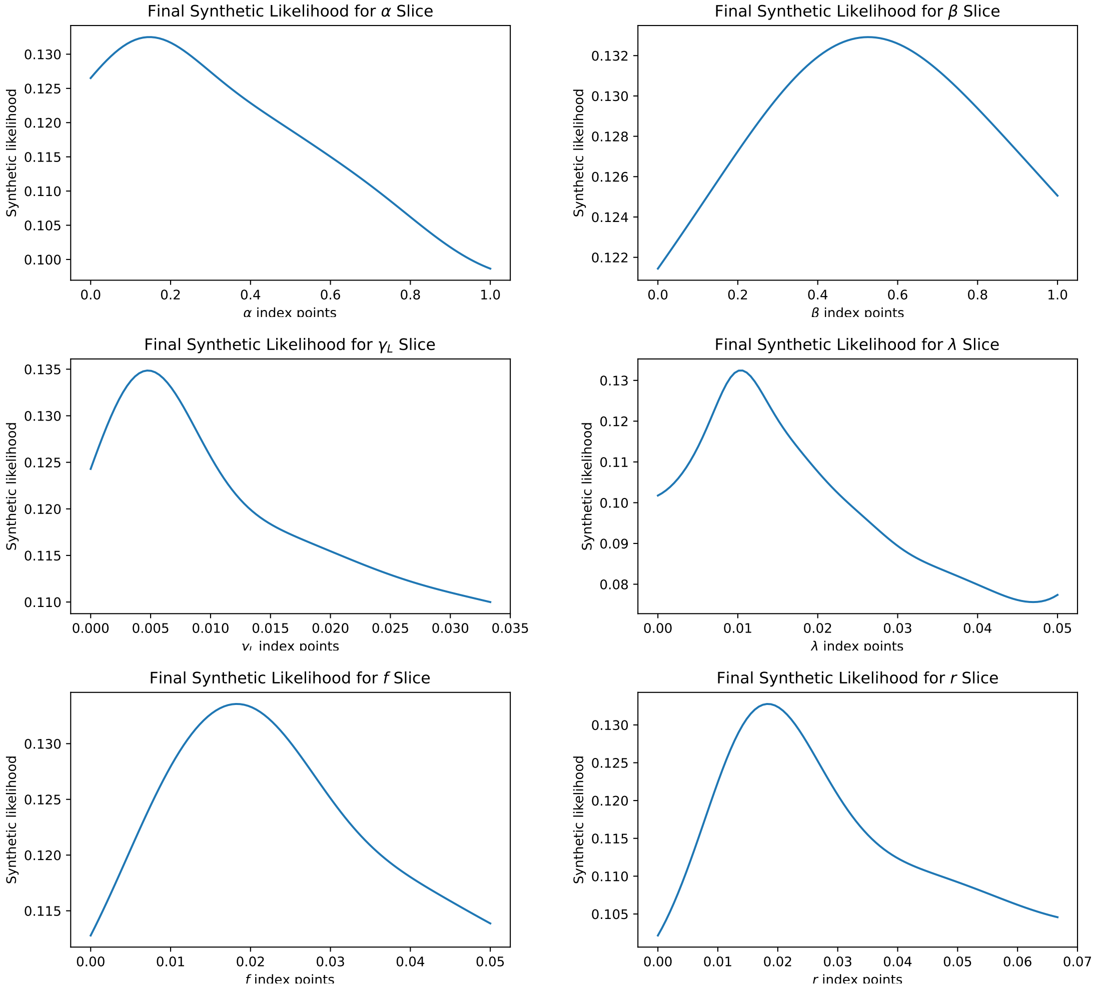


Figure 8.2: Final synthetic likelihoods $\hat{L}(\boldsymbol{\theta})$ after 500 iterations of Bayesian acquisition. Each likelihood shows how that likelihood changes across the parameter space.

8.2 Discussion

Although most of the procedure we used closely follows the paper by Gutmann and Cor 2016, there are a few key ways in which we modified the manuscript's method. In particular, we felt the choice of squared exponential kernel was not a good assumption, since it implicitly assumes a high degree of smoothness in the target discrepancy function. This is unlikely to be met if the model has any bifurcation points. Under the squared exponential function, if the true function has any sharp declines, or general non-smoothness , the length scale is forced to be set very small. Therefore, although the squared exponential is the most commonly chosen kernel, we used a Matérn kernel with $\nu = 5/2$. This is not as constrained as a squared exponential kernel, but realisations are twice mean square differentiable.

Gutmann and Cor 2016 use the lower confidence bound acquisition function, where the exploration parameter is written as $\eta_t := \sqrt{2\ln(\frac{t^{2d+2}\pi^2}{3\varepsilon})}$. This seems to be a mistake inherited from Brochu, Cora, and Freitas 2010. Even using Brochu's original citation (Srinivas et al. 2010) to revise this to $\eta_t := \sqrt{2\ln(\frac{t^{2d+2}\pi^2}{3\varepsilon})}$. [1] When we tried both of these exploration parameters, the choice of ε between $(0, 1)$ largely lead to repeated sampling from the same set of parameters, even for very small ε.

This is problematic, as Srinivas et al. 2010 assume a compact parameter space, whereas Gutmann and Cor 2016 do not.

They use a quadratic mean structure on unbounded parameters. This could be problematic in situations where the discrepancy has multiple local minima. When we tried it, the quadratic mean function poorly it resulted in in the acquisition function trying to sample extremely high parameter values. Therefore we bounded our parameters, making out parameter space compact.

8.3 Further Work

Moment matching could be used to avoid assuming that the distribution of the discrepancy $\mathcal{D}(\boldsymbol{\theta})$ or $\ln \mathcal{D}(\boldsymbol{\theta})$ is normal for fixed $\mathcal{D}(\boldsymbol{\theta})$. The sample variance could be explicitly modelled using

---

[1] One Python package that implements BOLFI notes this error, see: `https://github.com/elfi-dev/elfi/blob/dev/elfi/methods/bo/acquisition.py`

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.

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.

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.

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.

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` (visited on 03/26/2024).

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.

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`.

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`.

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` (visited on 03/26/2024).