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

List of Tables

List of Figures

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:

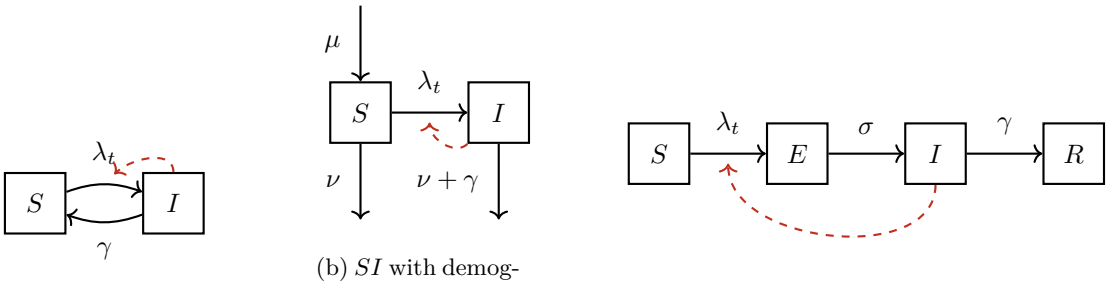
- S - Susceptable: at risk of contracting the disease
- E - Exposed: contracted the disease but not yet transmitting it
- I - Infectious (also called Infected): at risk of transmitting the disease
- R - Recovered: neither at risk of contracting or transmitting the disease.

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

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

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

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

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

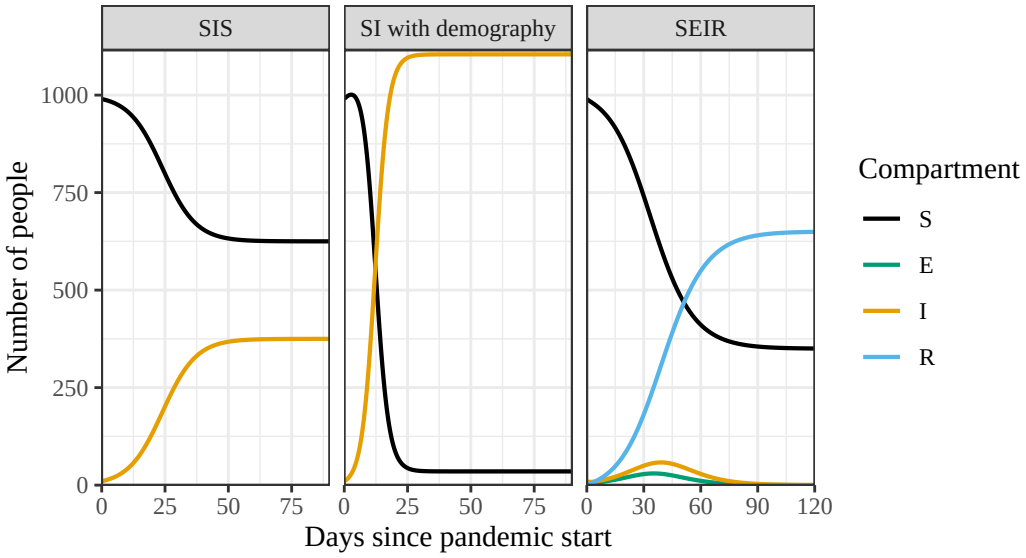


Figure 2.2: Solutions to the ordinary differential equations describing the models depicted in Figure 2.1. The initial infectious population was $I_0 = 10$, with $S_0 = 990$. In the $SEIR$ model, $E_0 = R_0 = 0$. For all models $\beta = 0.4$. For the SIS and SI model with demography $\gamma = 1/4$. For 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. $\mathcal{N}(t_1) - \mathcal{N}(t_0), \mathcal{N}(t_2) - \mathcal{N}(t_1), \ldots, \mathcal{N}(t_n) - \mathcal{N}(t_{n-1})$ *are independent for* $0 \leq t_0 < t_1 < \cdots < t_{n-1} < t_n$
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^*.

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

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 $\frac{\mathrm{d}\mathbb{E}(\mathcal{N}_2(t^*))}{\mathrm{d}t^*} = \gamma I_{t^*}$. Combining the two processes, we can see that the rate of change in the average number of people in S is

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

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 $\{\mathbf{C}_t\}_{t\geq 0} = \{C_1(t), C_2(t), \ldots, C_n(t)\}_{t\geq 0}$ where $C_i : \mathbb{R} \to \mathbb{N} \cup \{0\}$, is the number of people in compartment C_i, and for any fixed t, $\{C_1(t), C_2(t), \ldots, C_n(t)\}$ 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 ith 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, $\tau_i \sim \text{Exp}(\lambda_i)$. Therefore

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

□

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 $\mathbf{C}(0) := \{C_1(0), C_2(0), \ldots, C_n(0)\}$
while termination condition not met **do**
  Calculate intensities λ_i for all possible events $\mathcal{E}_i$
  Calculate total intensity $\lambda = \sum_i \lambda_i$
  Generate $\Delta t \sim \text{Exp}(\lambda)$
  Choose event 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.

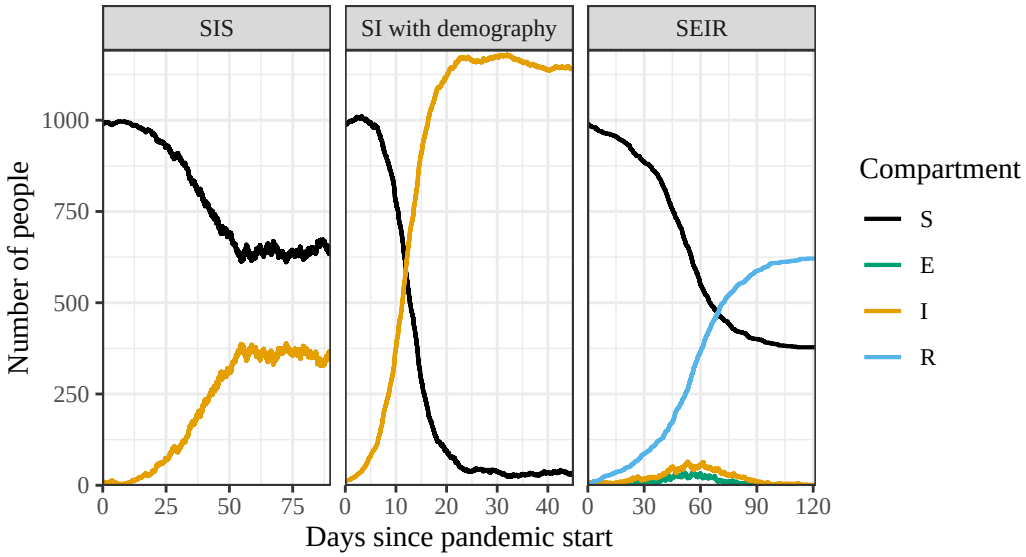


Figure 2.3: Exact stochastic simulations of the 3 different models using Algorithm 1. The parameters used were identical to those in Figure 2.2

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 λ_i for all possible events $\mathcal{E}_i$
  Calculate total intensity $\lambda = \sum_i \lambda_i$
  Choose a suitable time step τ (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

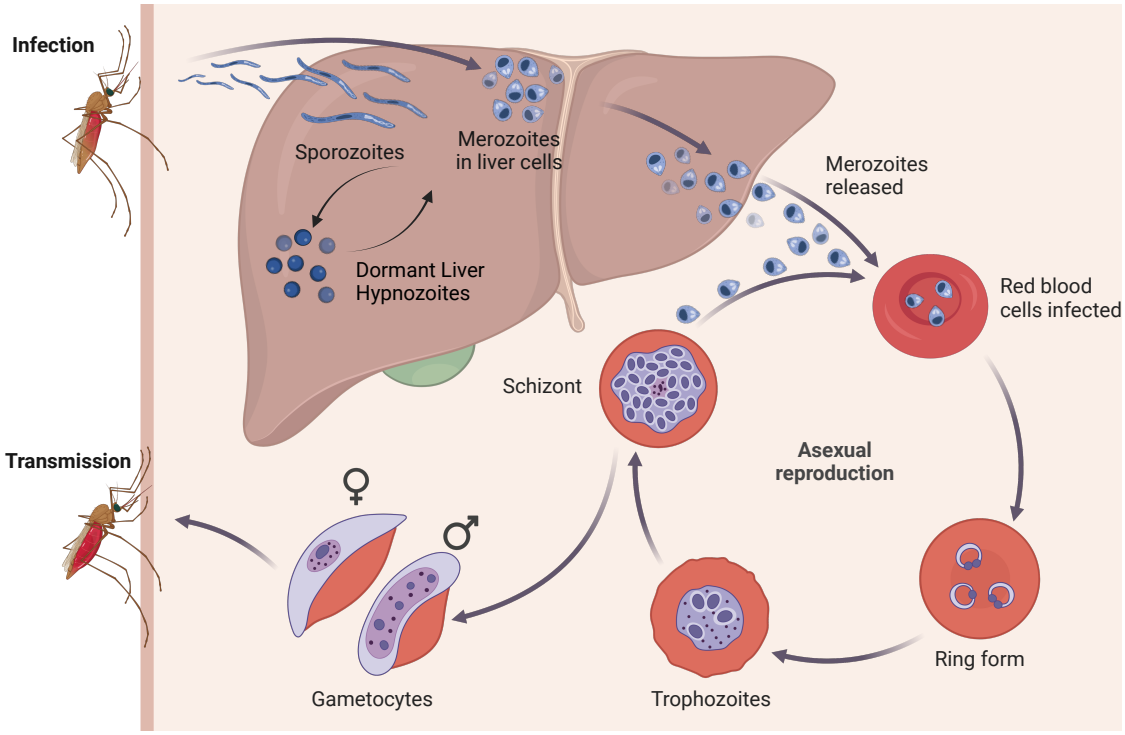


Figure 3.1: The *P. vivax* (malaria) lifecycle. *P. falciparum* does not have a dormant liver hypnozoite stage. Created with BioRender.com.

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

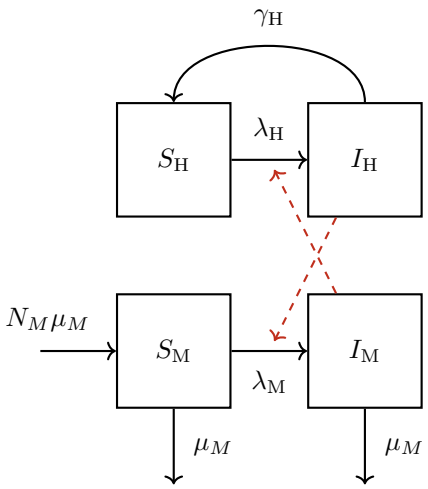


Figure 3.2: A simple Ross-Macdonald malaria model schematic, as described by Aron and May 1982. S_{H} and I_{H} are the number of susceptable and infected humans respectively, and S_{M} and I_{M} are the number of susceptable and infected mosquitos. The rate of human infection (λ_{H}) is dependant on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

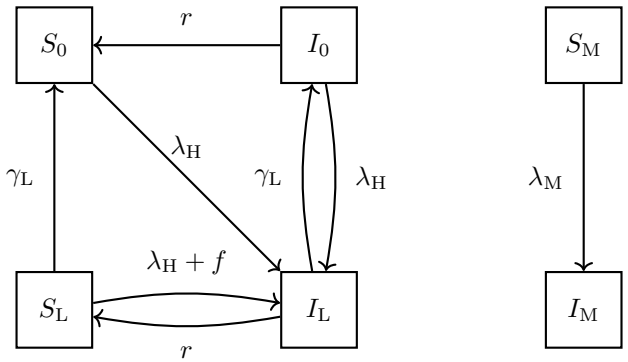


Figure 3.3: Diagram for *P. vivax* model in a tropical setting described by White et al. 2016. S and I are the number of susceptable and infected humans and mosquitos (denoted by subscript M). $\lambda_H = mabI_M$ and $\lambda_M = ac(I_0 + I_L)$

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:

$$
\begin{aligned}
\frac{\mathrm{d}S_0}{\mathrm{d}t} &= -\lambda_H S_0 + rI_0 + \gamma_L S_L \\
\frac{\mathrm{d}I_0}{\mathrm{d}t} &= -\lambda_H I_0 - rI_0 + \gamma_L I_L \\
\frac{\mathrm{d}S_L}{\mathrm{d}t} &= -\lambda_H S_L + rI_L - fS_L - \gamma_L S_L \\
\frac{\mathrm{d}I_L}{\mathrm{d}t} &= \lambda_H(S_0 + I_0 + S_L) - rI_L + fS_L - \gamma_L I_L \\
\frac{\mathrm{d}S_M}{\mathrm{d}t} &= g - \lambda_M(pS_M - (1-p)I_M) - gS_M \\
\frac{\mathrm{d}I_M}{\mathrm{d}t} &= \lambda_M(pS_M - (1-p)I_M) - gI_M. \qquad (I_0 + I_L = \text{total number of bloodstage infections})
\end{aligned}
$$

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

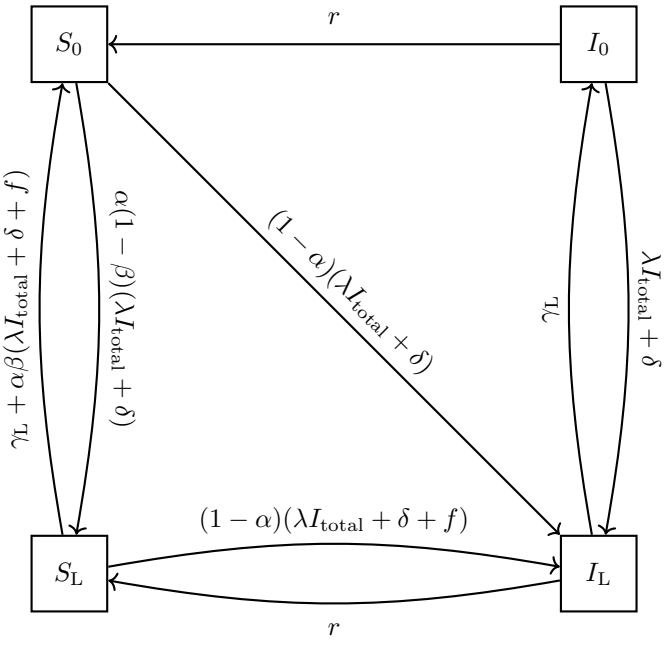


Figure 3.4: Diagram for *P. vivax* model described by Champagne et al. 2022. $I_\mathrm{total} = I_0 + I_\mathrm{L}$. Since the mosquito dynamics have been removed, λ now not has no dependencies on the number of infectious mosquitos.

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

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

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 $\mathbf{y}^{\text{obs}} = (y_1^{\text{obs}}, \ldots, y_n^{\text{obs}})$. Consider $\mathbf{y}(\boldsymbol{\theta}) = (y_1(\boldsymbol{\theta}), \ldots, y_n(\boldsymbol{\theta}))$ some model simulation of $\mathbf{y}^{\text{obs}}$. Often the observed data has some underlying index set $x_1, \ldots, x_n$, where x_i might be something like time. In this case we can also consider the observed data to be $\{(x_1, y_1^{\text{obs}}), \ldots, (x_n, y_n^{\text{obs}})\}$, and the model simulated data to be $\{(x_1, y_1(\boldsymbol{\theta})), \ldots, (x_n, y_n(\boldsymbol{\theta}))\}$.

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*

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

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

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

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*

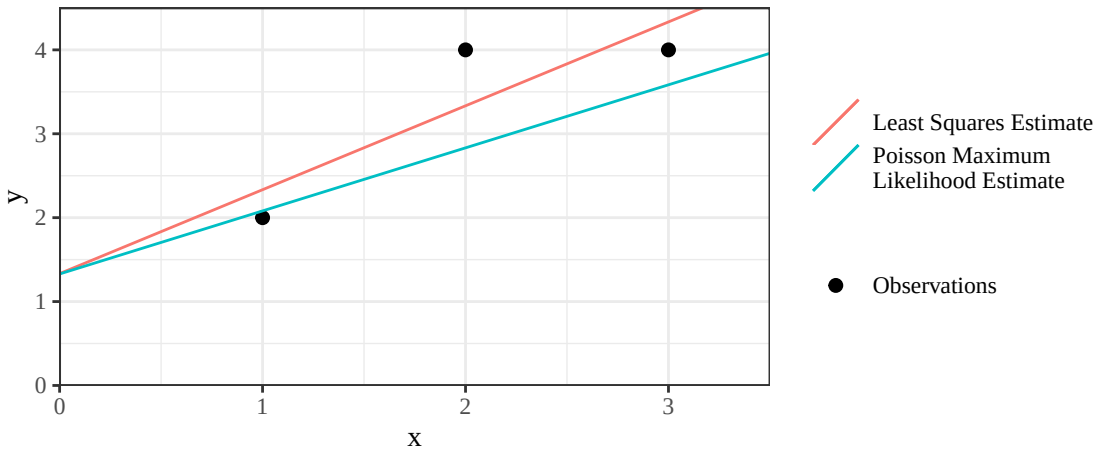


Figure 4.1: Two linear models of the form $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ fit given the set of observations $\{(1,2),(2,4),(3,4)\}$ using the method of least squares and maximum likelihood under the assumption that the data are independent realisations by a Poisson distribution with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

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* $\mathbf{y}^{\text{obs}}$ *fixed, the* likelihood function *is*

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

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

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

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

A natural estimate for $\boldsymbol{\theta}$ is the one that maximises the likelihood function $\mathcal{L}$, as it coincides with the value of $\boldsymbol{\theta}$ 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* $y_i(\boldsymbol{\theta}) + \varepsilon_i \sim \text{Pois}(y_i(\boldsymbol{\theta}))$*, where* $y_i(\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ *as previously defined. Therefore the maximimum likelihood estimate of* $\boldsymbol{\theta}$ *is*

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

which we 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*

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

Proof.

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

□

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 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

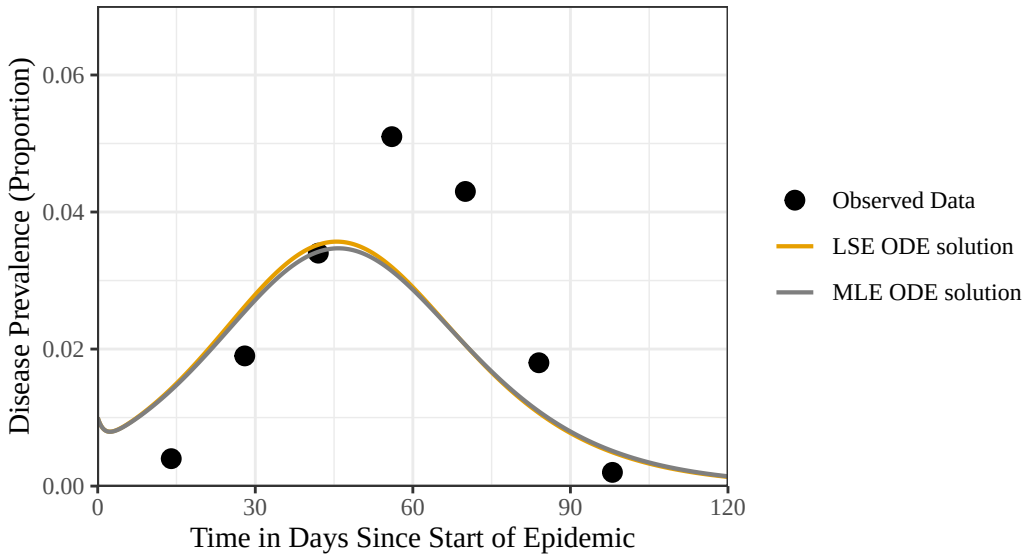


Figure 4.2: An $SEIR$ model fit to some observed prevalence data taken every two weeks over a 14 week period, genetated as I_t/N from the $SEIR$ simulation in Figure 2.3. All parameters were considered known except for β. The least squares estimate (LSE) $\beta^{\text{LSE}} = 0.3516$ for β was found 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 $U \sim \text{Unif}(0,1)$
if $U \leq \frac{g(X^*)}{Mp(X^*)}$ **then**
  return X^* as a sample from the distribution of X
else
  Repeat
end if

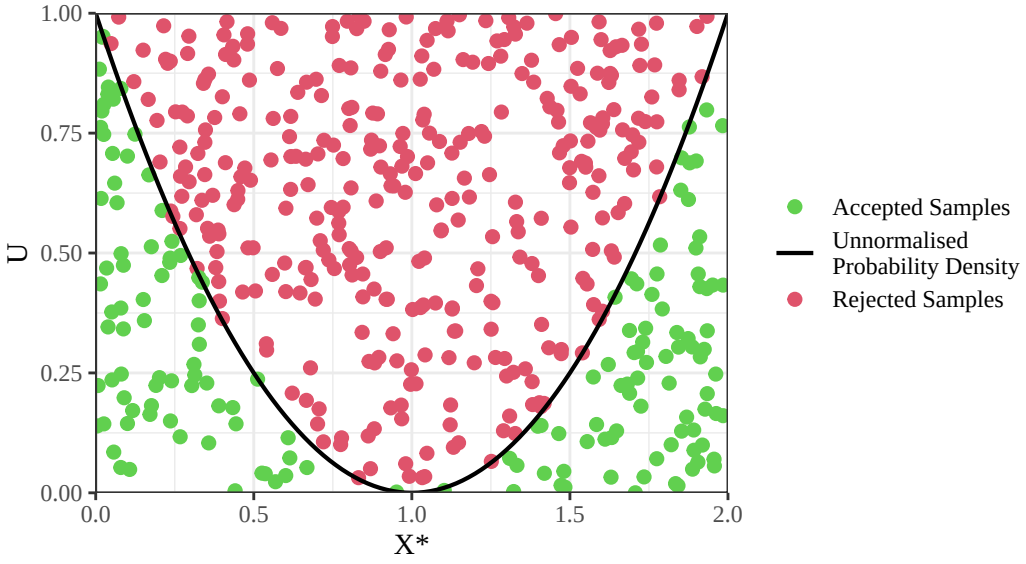


Figure 4.3: Samples of X from the unnormalised density $g(x) = (x-1)^2$ with $x \in (0,2)$ using the rejection sampler. $X^* \sim \text{Unif}(0,2)$ and $M = 1$. Green dots are samples from from X. Of 500 samples of X^*, 157 were accepted as samples of X.

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

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 $\text{Unif}(0,2)$ random variable. Therefore to generate samples from g we sample uniformly from $X^* \sim \text{Unif}(0,2)$, and then accept the sample if a new $U \sim \text{Unif}(0,1)$ 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 $\Pr(\boldsymbol{\theta}|\mathbf{y}^{\text{obs}})$ 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 distri-

bution $p(x)$. 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 $\mathcal{X} = \{1, 2\}$.

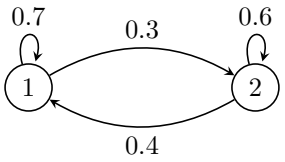


Figure 4.4: A simple time homogeneous Markov chain, with two states. It is characterised by the transition kernel $K(1,1) = \Pr(X_{i+1} = 1|X_i = 1) = 0.7$, $K(1,2) = \Pr(X_{i+1} = 2|X_i = 1) = 0.3$, $K(2,1) = \Pr(X_{i+1} = 1|X_i = 2) = 0.4$, and $K(2,2) = \Pr(X_{i+1} = 2|X_i = 2) = 0.6$. The stationary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

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

□

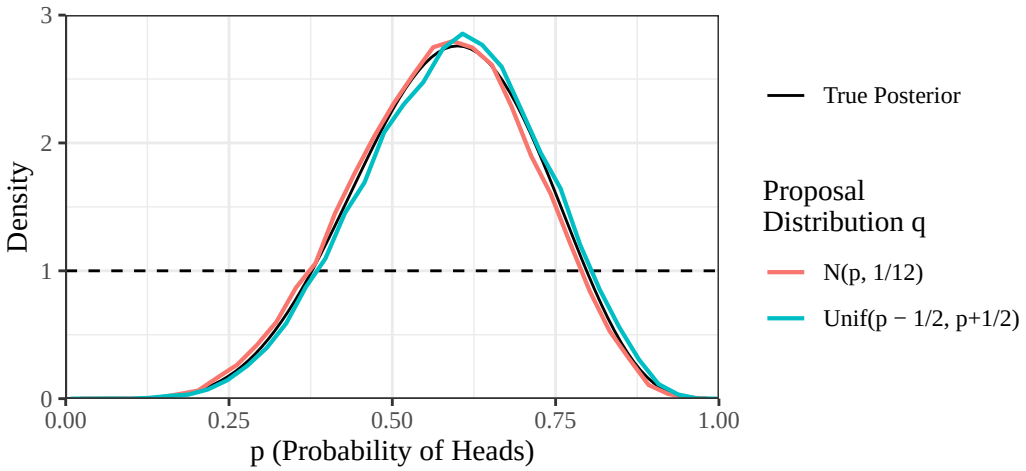


Figure 4.5: Samples from the posterior distribution of p using the Metropolis-Hastings algorithm. p was assumed to have a uniform prior between 0 and 1, with $y^{\text{obs}} = 6$, generated from Binom$(10, p)$, The choice of proposal distribution did not impact the final estimate of $\Pr(p|y^{\text{obs}})$.

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 $U \sim$ *Uniform*$(0,1)$
  if $U \leq \alpha$ ***then***
    $p_i \leftarrow X'$
  else
    $p_i \leftarrow x_{i-1}$
  end if
end for

return $\{p_0, p_1, \ldots, p_N\}$

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).

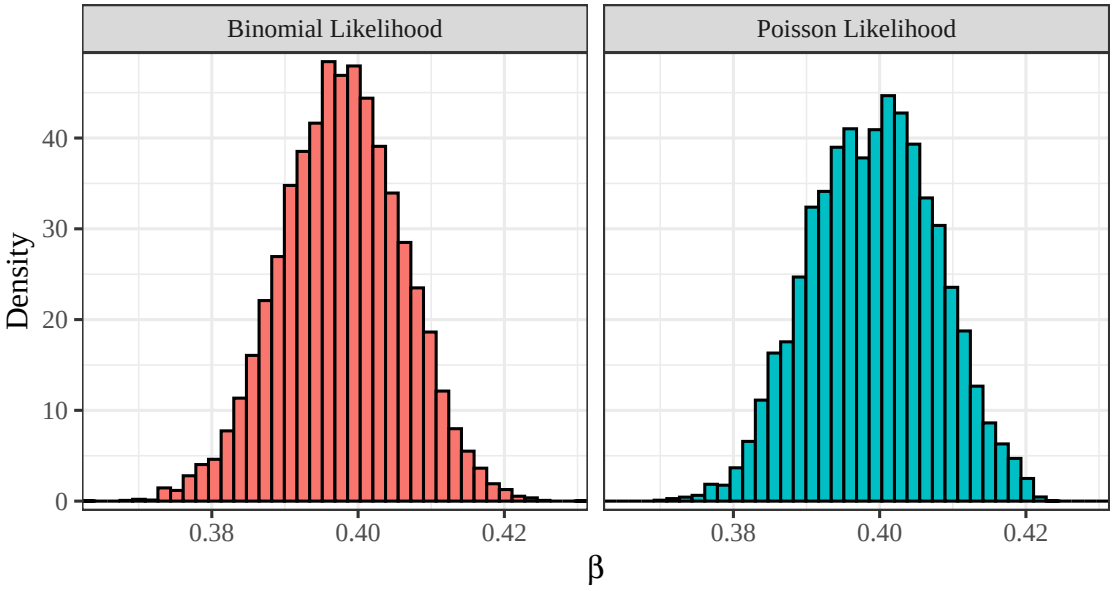


Figure 4.6: Given a daily incidence of $y^{\text{obs}} = 26$ at day 30 of an SIS epidemic, with unknown β, we use Metropolis-Hasting to sample from $\Pr(\beta|y^{\text{obs}})$. $\gamma = 1/4$ was assumed to be correct, and we compared the assumption $y^{\text{obs}} \sim \text{Binom}(\lfloor S_{30} \rfloor, \beta I_{30}/N)$, 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 $\mathbb{E}(\beta) = 1/3$. 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 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})$
for $i = 1$ to N **do**
  Sample $\theta_1^{(i)} \sim \Pr(\theta_1 | \theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)})$
  Sample $\theta_2^{(i)} \sim \Pr(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)})$
  $\vdots$
  Sample $\theta_d^{(i)} \sim \Pr(\theta_d | \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{d-1}^{(i)})$
  Save $(\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_d^{(i)})$ as $\boldsymbol{\theta}^{(i)}$
end for
return $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}\}$

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*

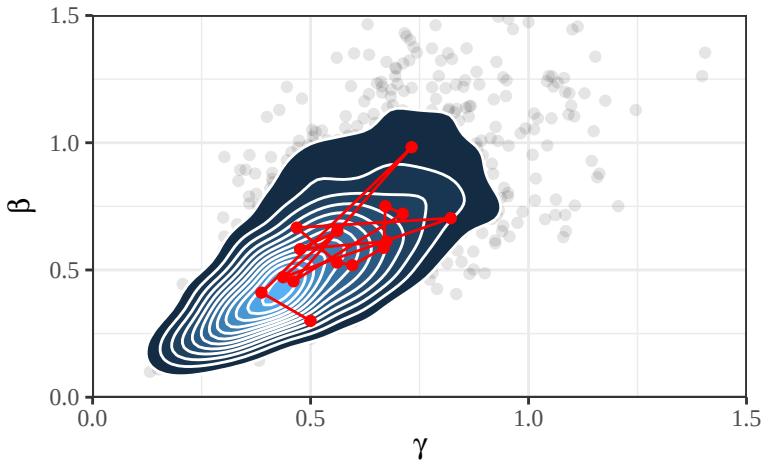


Figure 4.7: 2000 posterior samples from $\Pr(\beta, \gamma|\mathbf{y}^{\text{obs}})$, where $\beta|\gamma, \mathbf{y}^{\text{obs}} \sim \text{Gamma}(9, 4/\gamma + 4 + 8\gamma)$ and $\gamma|\beta, \mathbf{y}^{\text{obs}} \sim \text{InvGamma}(12, 12\beta)$. The samples were obtained using a Gibbs sampler.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

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

5.1 Gaussian Processes

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

Definition 5.1 (Gaussian Process)**.** *A collection of random variables $\{f(x)\}_{x\in\mathcal{X}}$ (where x may be a vector) is a* Gaussian process *if any finite subset of the collection of random variables is multivariate normal distributed. That is, there is a function $m : \mathcal{X} \to \mathbb{R}$ 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 $f(\mathbf{x}) := [f(x_1), f(x_2), \ldots, f(x_n)]^T$*

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

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

are

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

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. The most famous example of a Gaussian process is Brownian motion.

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

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

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

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

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

Definition 5.5 (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*

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

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

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

Kernels

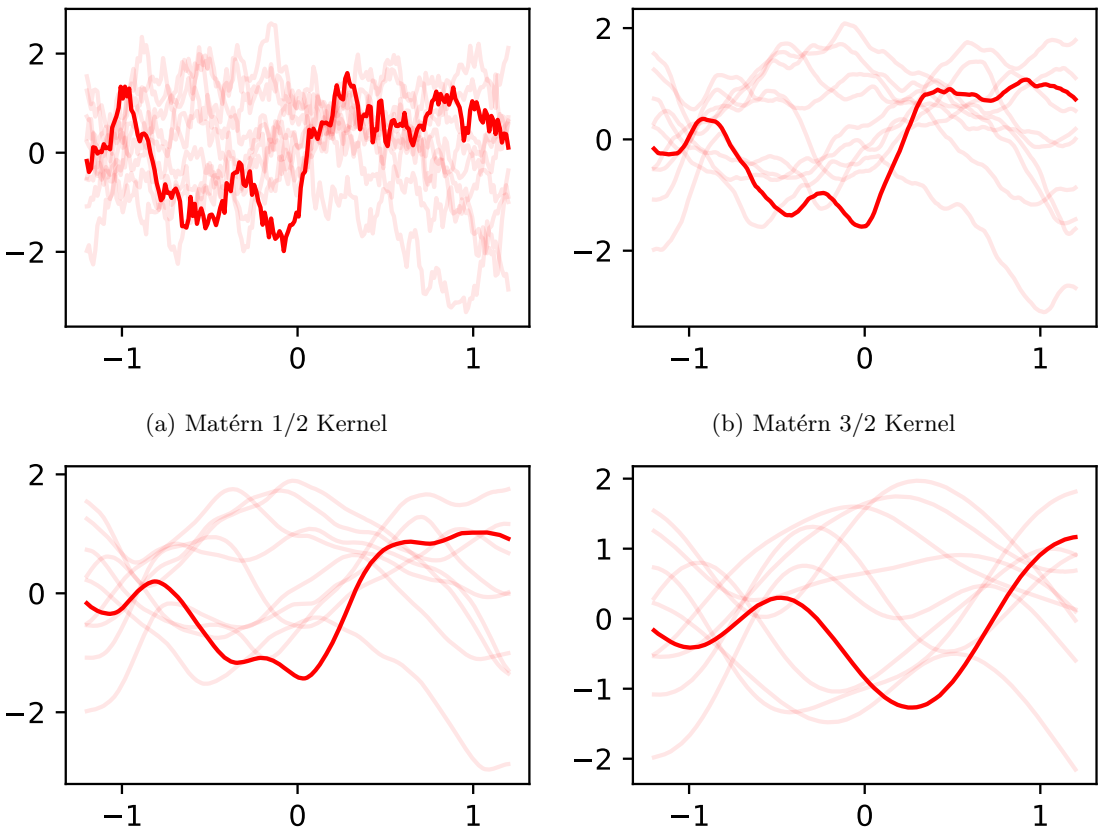
Matérn Kernel Family

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

Definition 5.7 (Matèrn Kernel)**.**

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

where K_ν is a modified Bessel function (defined in Abramowitz and Stegun 2013, p. 374). ν, ℓ, and σ_k are hyperparameters.



(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 one bolded. Samples for each kernel were generated from the same seed and the hyperparameters ℓ, and σ_k were set to 1.

Realisations from zero mean Gaussian processes with this kernel are $\lfloor \nu \rfloor$ times mean square differentiable (Rasmussen and Williams 2008). The most common values for ν are $1/2, 3/2$ and $5/2$, which result in functions that are 0, 1, and 2 times mean square differentiable. In these cases

the kernel can be slightly simplified to:

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

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

and

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

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

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

Squared Exponential Kernel

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

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

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

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

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

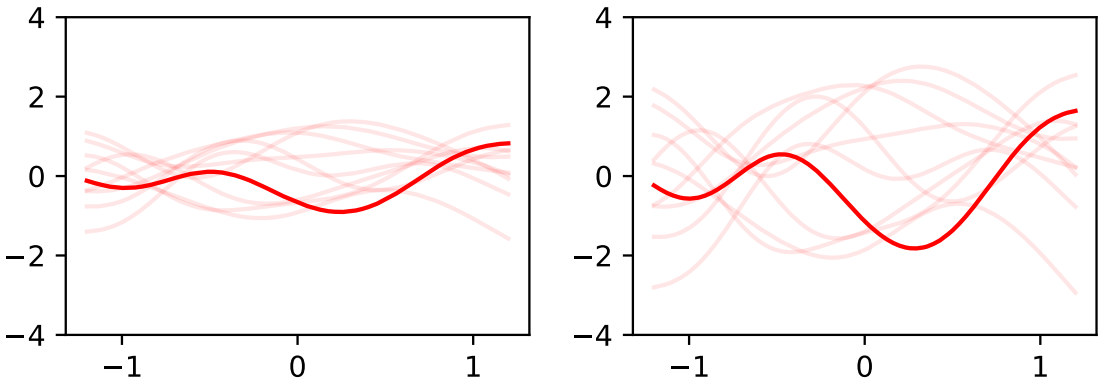
Length and Amplitude Hyperparameters

Both the Matérn and squared quadratic kernels (as well as most other common kernels choices), there are two hyperparameters 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.

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

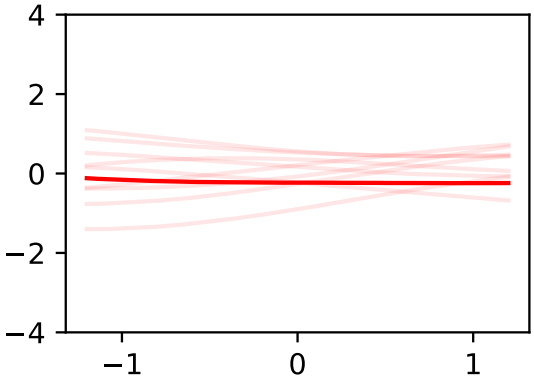
Definition 5.9 (Symmetric)**.** *A (square) matrix* $\mathbf{A}$ *is* symmetric *if* $A = A^T$

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

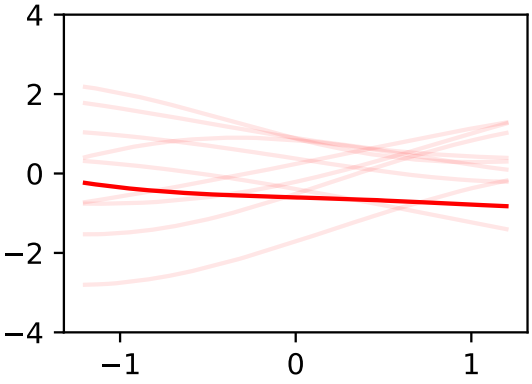


(a) $\ell = \sigma_k^2 = 1/2$

(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

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

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

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

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

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

5.2 Gaussian Process Regression

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

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

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

then

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

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

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

by the inverse of a block matrix. Therefore

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

Since

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

$f(\mathbf{x}) - m(\mathbf{x})|f(\mathbf{x}_*)$ 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,

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

□

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

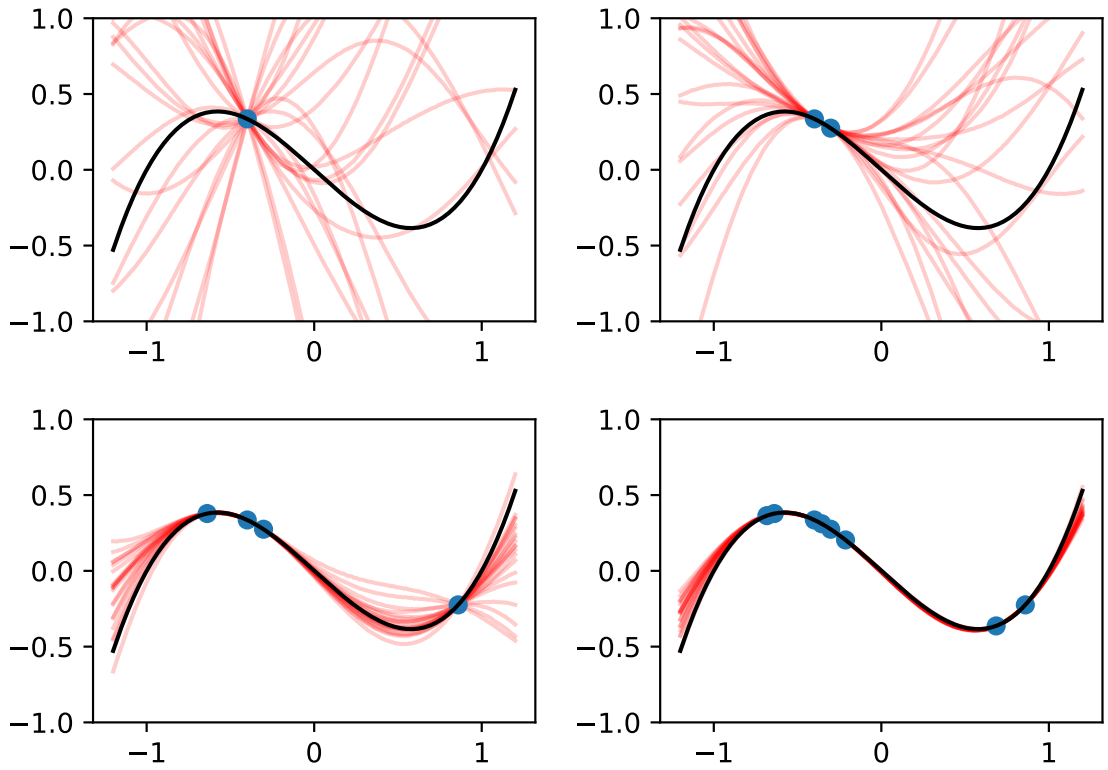


Figure 5.3: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 1, 2, 4, and 8 observations in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was zero mean and had a squared exponential kernel. The hyperparameters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

$f(x)$. The more points that the Gaussian process is conditioned on, the more certainty there is in the sample paths, seen in Figure 5.3.

Observation Variance

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

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

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

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

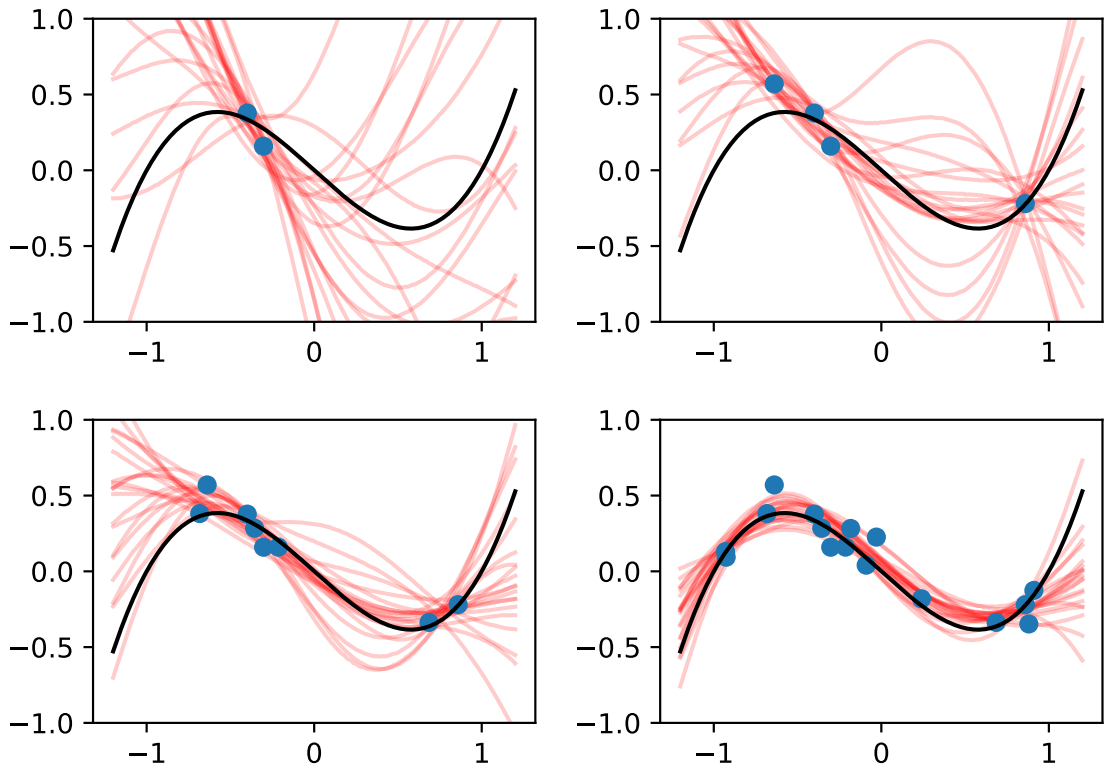


Figure 5.4: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 2, 4, 8, and 16 observations of $f(x_i)+\varepsilon_i$, where ε_i is i.i.d. $\text{MVN}(0, \sigma_o^2)$ with $\sigma_o^2 = 0.01$ in blue. The red lines are new realisations from the conditioned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

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

5.3 Model Selection

Kernel

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

Hyperparameters

The kernel hyperparameters ℓ and σ_k^2 are generally not fixed *a prior*. Similarly, the observation variance σ_o^2 hyperparameter may not be known. There are two main (frequentist) ways to fit these hyperparameters: maximum likelihood estimation, and leave-one-out cross validation.

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

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

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.

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

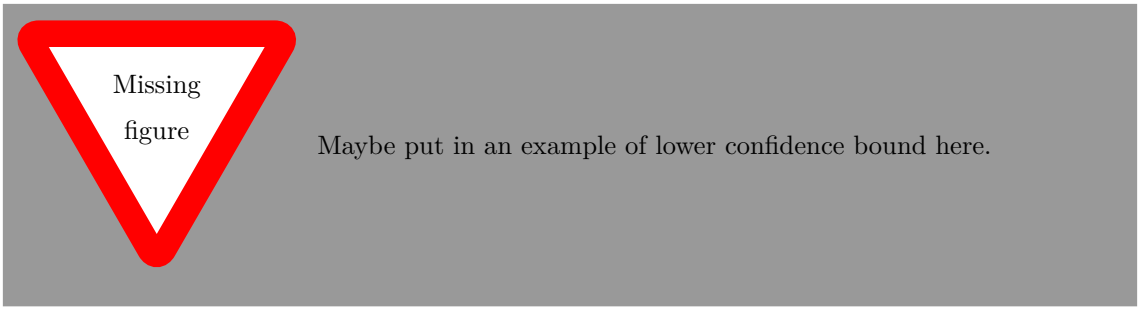
where $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ is the distribution of the ith element of $f(\mathbf{x}_*)$ conditioned on the rest of the observed data excluding that element (represented by $f_{-i}(\mathbf{x}_*)$). $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ can be found by Theorem 5.13. Computationally 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).

Recent work has shown that at least under specific conditions, the leave-one-out estimates for the scale hyperparameter are more robust to a larger family of target functions (Naslidnyk et al. 2024), and the broader literature seems to favor leave-one-out cross validation (for example see Gutmann and Cor 2016).

Finally there is scope for a Bayesian approach to model selection. By setting priors on the hyperparameters (sometimes called hyper-priors) and using the likelihood as described in the maximum likelihood estimation approach, a posterior distribution can be easily contrived. Samples



could then be taken from the posterior density of the hyperparameters. Alternatively a point estimate could be taken by choosing the maximum a posteriori point. This approach has some obvious benefits, particularly when taking a posterior sample, which can capture some of the uncertainty in the model more so than a point estimate.

5.4 Bayesian Acquisition Functions

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

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

Lower Confidence Bound

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

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

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

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

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

and are highly dependent on choice of covariance function and dimensionality of the parameter space.

Probability of Improvement

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

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

$$\mathcal{A}_{\mathrm{PI}}(x) := \Pr(f(x) < \mu_*)$$

where $\mu_* := \min_{x_* \in \mathbf{x}_*} f(x_*)$.

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

The probability of improvement can also be expressed as

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

which motivates the form of the next acquisition function.

Expected Improvement

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

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

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

where $\mu_* := \min_{x_* \in \mathbf{x}_*} f(x_*)$.

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

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

Part II

Calibrating Parameters for a *P. vivax* Model

Chapter 6

Methods

6.1 Creation of Synthetic Data

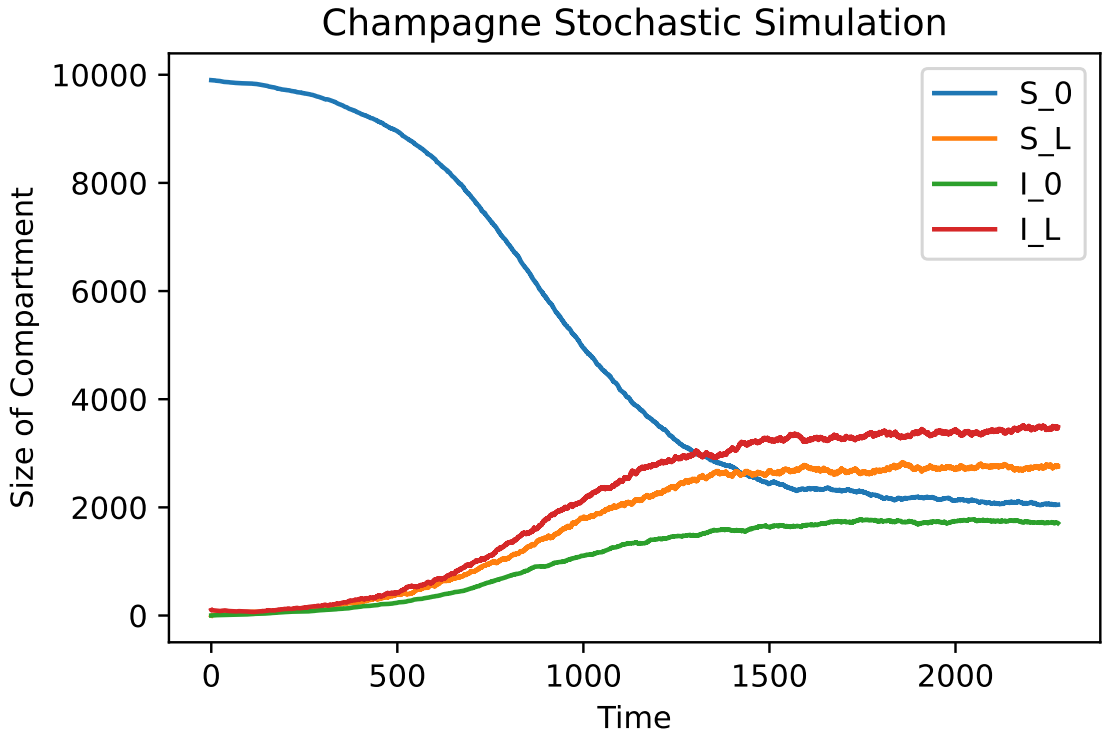


Figure 6.1: A Doob-Gillespie Simulation of the model described by Champagne et al. 2022 with $\alpha = 0.4$, $\beta = 0.4$, $\gamma_L = 1/223$, $\lambda = 0.04$, $f = 1/72$, $r = 1/60$, and $\delta = 0$. The population was 10000, with 100 initial infections (both blood and liver stage I_L).

We investigated the model by Champagne et al. 2022 as described in 3.2. A malaria epidemic was simulated using the Doob-Gillepsie algorithm shown in Figure 6.1, using a population size of 10,000, and initial infected population of 100 (with both liver and blood stage infection). The parameters used closely followed those reported in Champagne et al. 2022, with the exact parameters used reported in Table 6.1. $\delta = 0$ was assumed to be known. From intialisation, the simulation was run for 200,000 events (with an event being anything that caused the size of any compartment

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

Parameter description	Parameter	Value	Units
effective blood stage treatment proportion	α	0.1235	None
effective liver stage treatment proportion	β	0.429	None
rate of liver stage disease clearance	γ_L	1/383	1/days
rate of infection	λ	0.01	1/days
rate of relapse	f	1/69	1/days
rate of blood stage disease clearance	r	1/60	1/days.

to change such as an infection, recovery, relapse etc.), after which, the model was assumed to have reached steady state behaviour. A number of events was chosen because time to convergence is highly dependent on the scales of the parameters.

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

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

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. The observed data was taken to be from the simulated case counts (incidence) and prevalence (as the absolute number of people infected) of the simulated epidemic, described in Table 6.2.

6.2 Model Simulations and Discrepancy Function

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

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

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

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

6.3 Gaussian Process and Initialisation

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

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

where the $\ln \mathcal{D}_j(\boldsymbol{\theta})$s are 30 i.i.d. samples generated by model runs. These were run in parallel on a supercomputer. The samples $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$ were assumed to be noisy observations of $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})] + \varepsilon$, where $\varepsilon \sim N(0, \sigma_o^2)$, by the central limit theorem. σ_o^2 was assumed to be independent of $\boldsymbol{\theta}$, and has the natural intepretation as the variance of the sample mean. $d_{\mathcal{GP}}(\boldsymbol{\theta})$ was assume to have unknown constant mean $m_{\mathcal{GP}}$, and kernel

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

where

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

This kernel is, a Matérn kernel with $\nu = 5/2$ and automatic relevance determination - i.e. each parameter $\theta \in \boldsymbol{\theta}$ was scaled by ℓ_θ. In effect, this assigns each parameter its own length hyperparameter.

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

Parameter	Upper Bound	Unit
Proportion of treatment clearing blood stage disease 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

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

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

Hyperparameter	Description
σ_o^2	Observation variance (var($\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$))
σ_k^2	Matérn kernel amplitude
ℓ_α	Length parameter associate with α
ℓ_β	Length parameter associate with β
ℓ_{γ_L}	Length parameter associate with γ_L
ℓ_λ	Length parameter associate with λ
ℓ_f	Length parameter associate with f
ℓ_r	Length parameter associate with r
$m_{\mathcal{GP}}$	Gaussian process mean

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 6.3). For each set of parameters, $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$ was generated. The hyperparameters described in 6.4 were selected by leave one out cross validation, and $d_{\mathcal{GP}}^{(0)}(\boldsymbol{\theta})$ was fit to the samples.

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

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

6.4 Bayesian Acquisition and Parameter Updates

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

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

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

This was done because when $\mathcal{A}_{\text{EI}}(\boldsymbol{\theta})$ is very small, particularly if $d_{\mathcal{GP}}(\boldsymbol{\theta}^*)$ is large, the gradient of $\mathcal{A}$ can be negligible, causing the convergence criteria to be met prematurely. Thus, random initialization is not conducive to the algorithm's success. Conversely, initialising at the current minimum $\arg\min_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} d_{\mathcal{GP}}(\boldsymbol{\theta}^*)$ can limit the explored parameter space, as the current minimum may be located near a local minimum.

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

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

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

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

Since

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

we can express $\hat{L}$ as

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

The Gaussian process and Gaussian process regression was implemented using TensorFlow (Martín Abadi et al. 2015), and all code is available at `https://github.com/jaycrick/masters_project`.

Chapter 7

Results and Discussion

7.1 Validation

Table 7.1: Final Gaussian process hyperparameters

Hyperparameter	Final value
σ_o^2	0.07
σ_k^2	0.707
ℓ_α	0.324
ℓ_β	0.715
ℓ_{γ_L}	0.010
ℓ_λ	0.006
ℓ_f	0.016
ℓ_r	0.016
$m_{\mathcal{GP}}$	0.879

The hyperparameters, reported in Table 7.1, and trained using leave one out cross validation and expected information acquisition function both converge smoothly, as seen in Figures 7.2 and 7.3, suggesting the gradient descent algorithm is well optimised. The violin plot of discrepency function in Figure 7.1 is well dispersed suggesting a good amount of exploration has been done, but the majority of samples are in the low discrepency region. This suggests a good exploration exploitation trade-off.

Figure 7.4 suggests that after 500 iterations $d_{\mathcal{GP}}(\boldsymbol{\theta})$ fits to $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$ well. The λ slice has the sharpest minimum, whereas β does not appear to have much impact on the discrepancy. As the number of iterations increased, the Gaussian process visibly improved at fitting to the mean. Interum iterations of $d_{\mathcal{GP}}^{(t)}(\boldsymbol{\theta})$ can be seen in the appendix.

7.2 Parameter estimation

The synthetic likelihood of each parameter can be seen in Figure 7.6. Unsurprisingly, λ has the sharpest peak in the likelihood function. All likelihoods were unimodal. The maximum likelihood estimate for $\boldsymbol{\theta}$ as reported in Table 7.2 is very close to the true values of the parameters after 500 iterations. This suggests that the synthetic likelihood approximates the true likelihood well. The maximum likelihood slice estimates were the univariate maximum likelihood estimates holding all other parameters at the true value. These are the peaks in Figure 7.6. Since the maximum

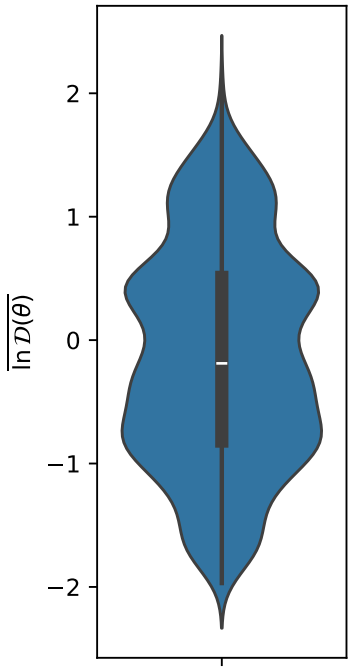


Figure 7.1: $\overline{\ln \mathcal{D}(\boldsymbol{\theta})}$ violin plot

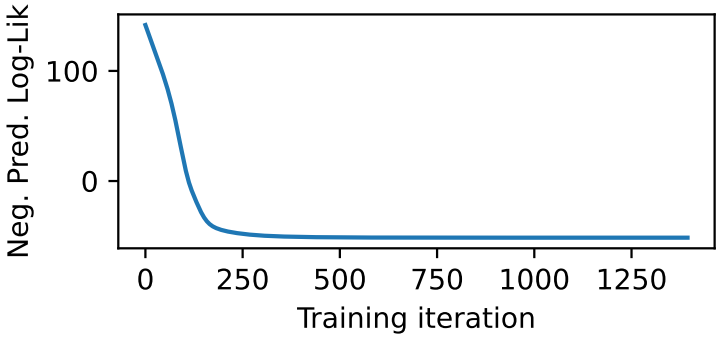


Figure 7.2: Hyperparameter training

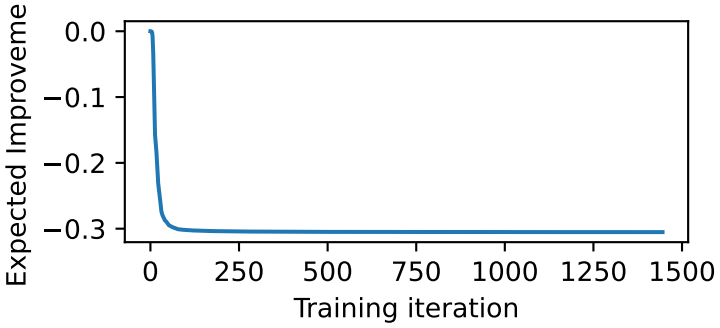


Figure 7.3: Finding $\arg\min_{\boldsymbol{\theta}} \mathcal{A}_{\mathrm{EI}}(\boldsymbol{\theta})$

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

Parameter	True	MLE	ML Slice Estimate
α	0.124	0.153	0.17
β	0.429	0.555	0.52
γ_L	0.0026	0.006	0.005
λ	0.01	0.01	0.001
f	0.014	0.024	0.02
r	0.017	0.023	0.02

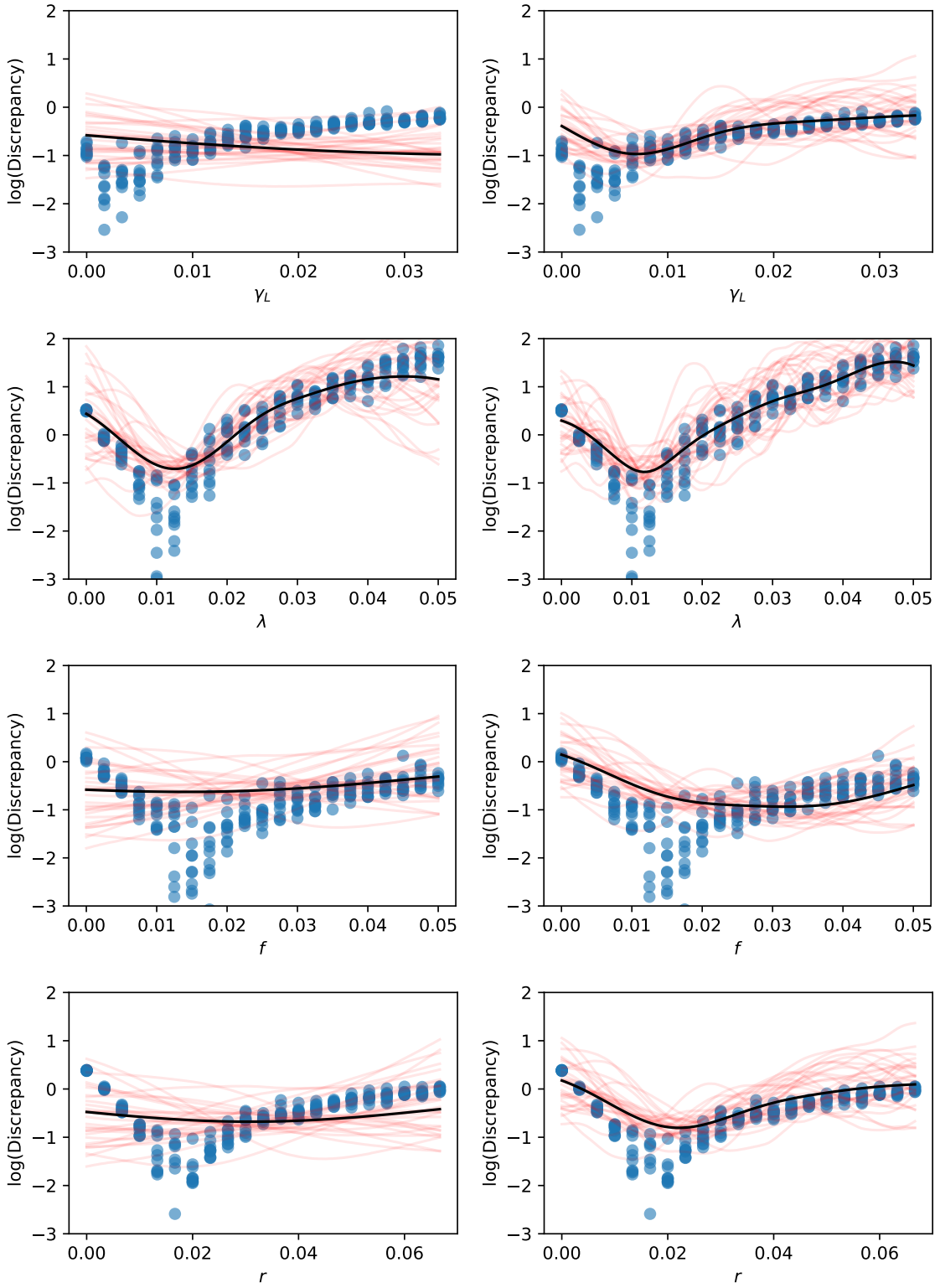


Figure 7.4: The left column of figures is the Gaussian process after initialisation $d^{(0)}_{\mathcal{GP}}(\boldsymbol{\theta})$. The black line is $\mathbb{E}(d^{(0)}_{\mathcal{GP}}(\boldsymbol{\theta}))$, and the red lines are multiple realisations of $d^{(0)}_{\mathcal{GP}}(\boldsymbol{\theta})$. The right column of figures is after 500 sampling iterations, with the black line being $\mathbb{E}(d^{(500)}_{\mathcal{GP}}(\boldsymbol{\theta}))$. The blue dots are realisations of $\ln \mathcal{D}(\boldsymbol{\theta})$, which $d_{\mathcal{GP}}$ approximates predict the mean of. The parameters are varied univariately, with all other parameters fixed at the true parameters.

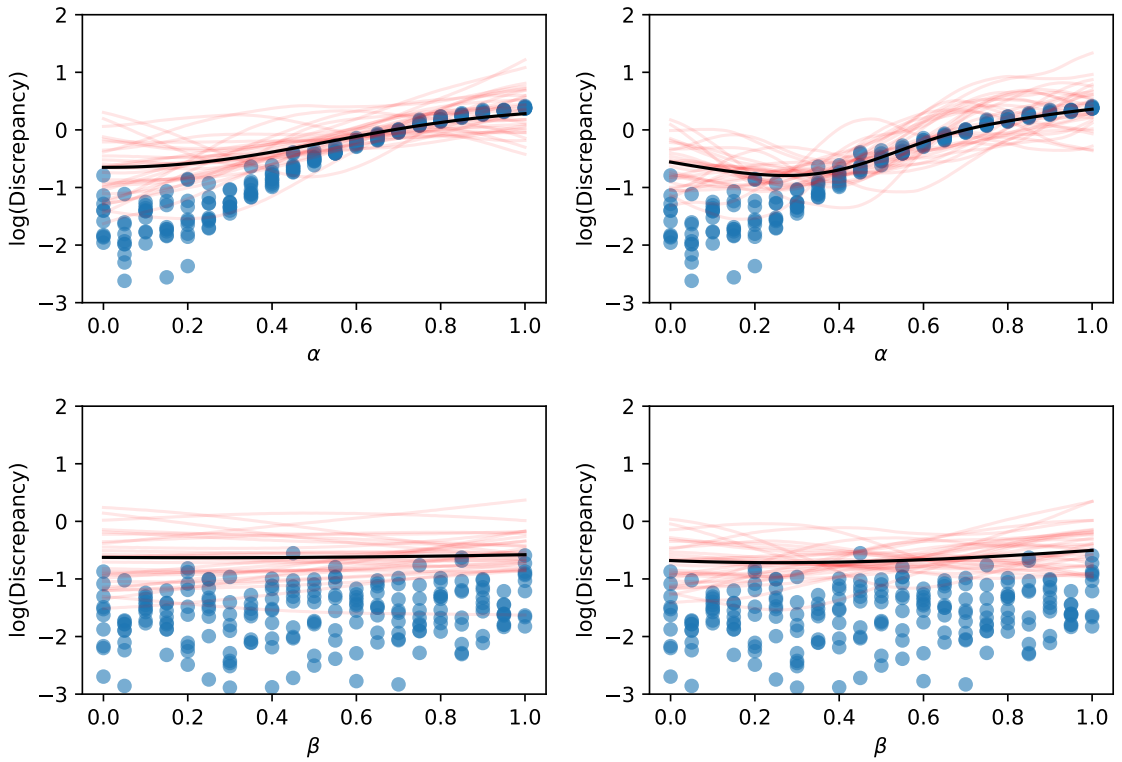


Figure 7.5: Gaussian process approximations of the treatment parameters, as with Figure 7.4

likelihood estimate was close to the true parameters, the maximum likelihood slice estimates were also close to the true values.

7.3 Discussion and Future Work

To our knowledge, the use of the synthetic likelihood as described above has not been used to calibrate a malaria model before. Champagne et al. calibrate the model we use by find the ODE equilibrium, and fitting a single parameter λ to incidence data. Not only were we able to effectively recover the true λ from a synthetic run, we were able to recover all model parameters simultaneously. A similar method has been set out for infectious diseases in Gutmann and Cor 2016.

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

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

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

Although the likelihood free procedure we outlined for this model closely resembles Gutmann

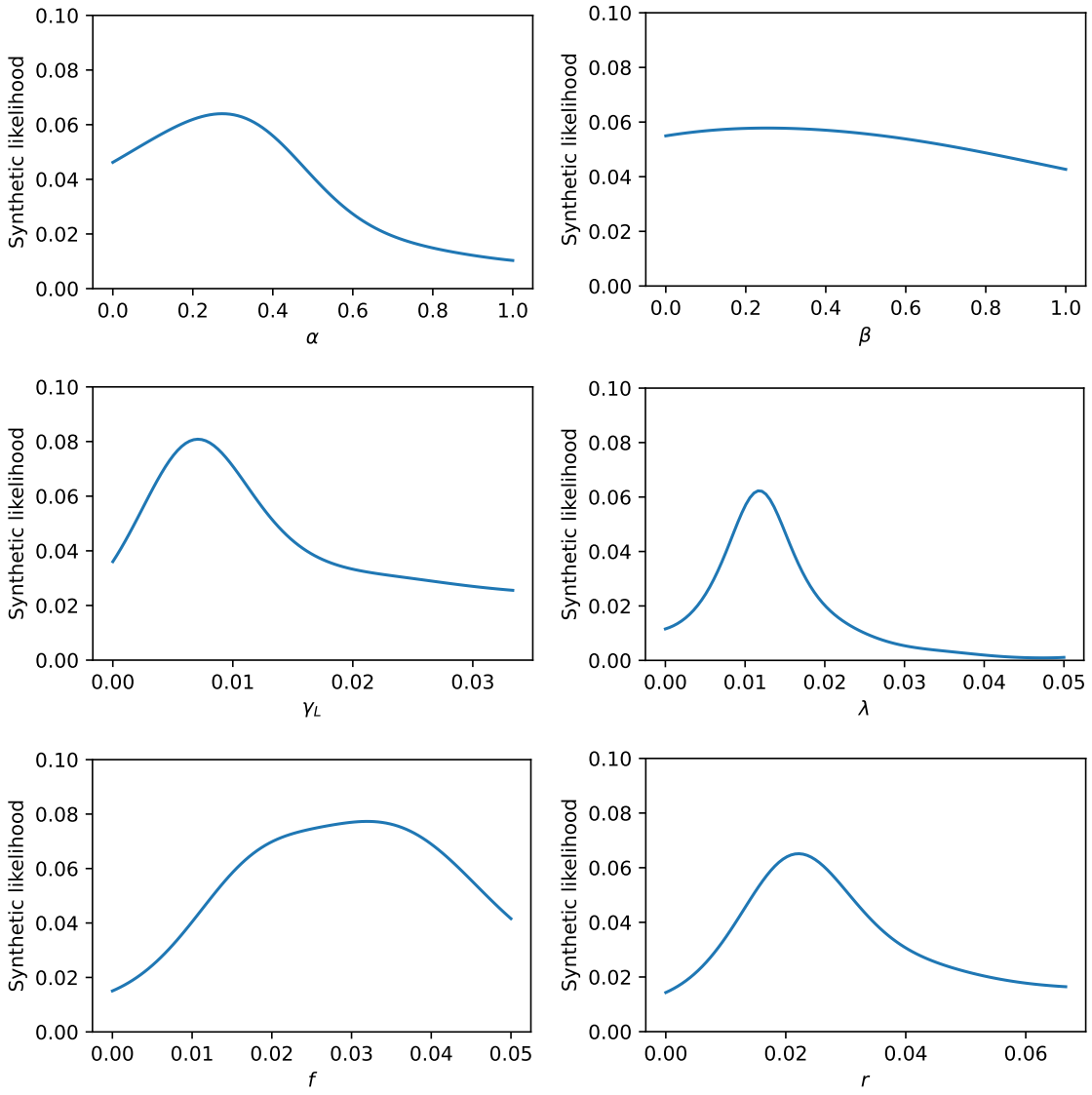


Figure 7.6: Final univariate synthetic likelihoods $\hat{L}(\boldsymbol{\theta})$ after 500 sampling iterations. All values not shown were fixed at the true parameters.

and Cor 2016, there are a few significant changes that improve on the method outlined in that manuscript.

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

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

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

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

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

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

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

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

This highlights another problem. Around bifurcation points, it is expected that $\mathbb{E}(\mathcal{D}(\boldsymbol{\theta}))$ behaves erratically. Gutmann and Cor 2016 use a squared exponential kernel for their Gaussian process approximation of cannot capture this behaviour without making any length scales very large. For an example, demonstrating the utility of the Matérn kernel over the squared exponential in the case of a non smooth function, see Jones 2021. Another possible solution is to use a

Student-t process to approximate the discrepancy, as it has heavier tails, so is more forgiving to sudden jumps. The multivariate Student-t distribution has some properties analogous to the multivariate normal distribution. This includes an analytic solution to the conditional distribution, similar to Theorem 5.13. For more details see Shah, Wilson, and Ghahramani 2014.

Gutmann and Cor 2016 use the lower confidence bound acquisition function, where the exploration parameter is the slowly increasing function

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

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

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

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

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

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

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

One way to reduce the computational overhead of this method would be to reduce the number of samples were used to calculate the sample mean. This paper used 30 to ensure convergence

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

to the normal distribution, however less could be taken with the trade off of a larger observation variance. The sample mean can be calculated in parallel, so the rate limiting step is miminising the acquisition function each iteration. Rather than sampling uniformly across a single parameter, multivariate noise could be added to $\boldsymbol{\theta}$ to sample multiple sample means at once. Resources should be maximally allocated to calculate as many $\mathcal{D}(\boldsymbol{\theta})$ as feasible in one step, split between multiple repeats for the sample mean, and multiple $\boldsymbol{\theta}$s for better training of the Gaussian process approximation.

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.

Fahrmeir, Ludwig, W. Hennevogl, and Gerhard Tutz (2013). *Multivariate Statistical Modelling Based on Generalized Linear Models*. OCLC: 1066189579. New York, NY: Springer. ISBN: 978-1-4899-0010-4.

Gani, Raymond and Steve Leach (Dec. 13, 2001). "Transmission potential of smallpox in contemporary populations". In: *Nature* 414.6865, pp. 748–751. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/414748a. URL: https://www.nature.com/articles/414748a (visited on 06/10/2024).

Gelman, Andrew et al. (2014). *Bayesian data analysis*. Third edition. Texts in statistical science series. Boca Raton London New York: CRC Press, Taylor and Francis Group. 667 pp. ISBN: 978-1-4398-4095-5.

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).

Jones, Andy (July 31, 2021). *The Matérn class of covariance functions*. Andy Jones. URL: https://andrewcharlesjones.github.io/journal/matern-kernels.html (visited on 06/18/2024).

Keeling, Matthew James and Pejman Rohani (2008). *Modeling infectious diseases in humans and animals*. OCLC: ocn163616681. Princeton: Princeton University Press. 366 pp. ISBN: 978-0-691-11617-4.

Martín Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Milner, Danny A. (Jan. 2018). "Malaria Pathogenesis". en. In: *Cold Spring Harbor Perspectives in Medicine* 8.1, a025569. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a025569. URL: http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a025569 (visited on 03/24/2023).

Naslidnyk, Masha et al. (2024). *Comparing Scale Parameter Estimators for Gaussian Process Interpolation with the Brownian Motion Prior: Leave-One-Out Cross Validation and Maximum Likelihood*. arXiv: 2307.07466 [math.ST].

Price, R.N. et al. (2020). "Plasmodium vivax in the Era of the Shrinking P. falciparum Map". English. In: *Trends in Parasitology* 36.6, pp. 560–570. ISSN: 1471-4922. DOI: 10.1016/j.pt.2020.03.009.

Rasmussen, Carl Edward and Christopher K. I. Williams (2008). *Gaussian processes for machine learning*. 3. print. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press. 248 pp. ISBN: 978-0-262-18253-9.

Robert, Christian P. and George Casella (2010). *Monte Carlo statistical methods*. 2. ed., softcover reprint of the hardcover 2. ed. 2004. Springer texts in statistics. New York, NY: Springer New York. 645 pp. ISBN: 978-1-4757-4145-2 978-1-4419-1939-7. DOI: 10.1007/978-1-4757-4145-2.

Shah, Amar, Andrew Gordon Wilson, and Zoubin Ghahramani (2014). *Student-t Processes as Alternatives to Gaussian Processes*. arXiv: 1402.4306.

Smith, David L. et al. (Apr. 2012). "Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens". en. In: *PLOS Pathogens* 8.4. Publisher: Public Library of Science, e1002588. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1002588. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002588 (visited on 03/28/2023).

Srinivas, Niranjan et al. (2010). "Gaussian process optimization in the bandit setting: no regret and experimental design". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, pp. 1015–1022. ISBN: 9781605589077.

White, Michael T. et al. (Mar. 30, 2016). "Variation in relapse frequency and the transmission potential of *Plasmodium vivax* malaria". In: *Proceedings of the Royal Society B: Biological Sciences* 283.1827, p. 20160048. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2016.0048.

URL: https://royalsocietypublishing.org/doi/10.1098/rspb.2016.0048 (visited on 08/22/2023).

Wikipedia contributors (2024). *Exponential family — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-May-2024]. URL: https://en.wikipedia.org/w/index.php?title=Exponential_family&oldid=1202463189.

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).

Chapter 8

Appendices

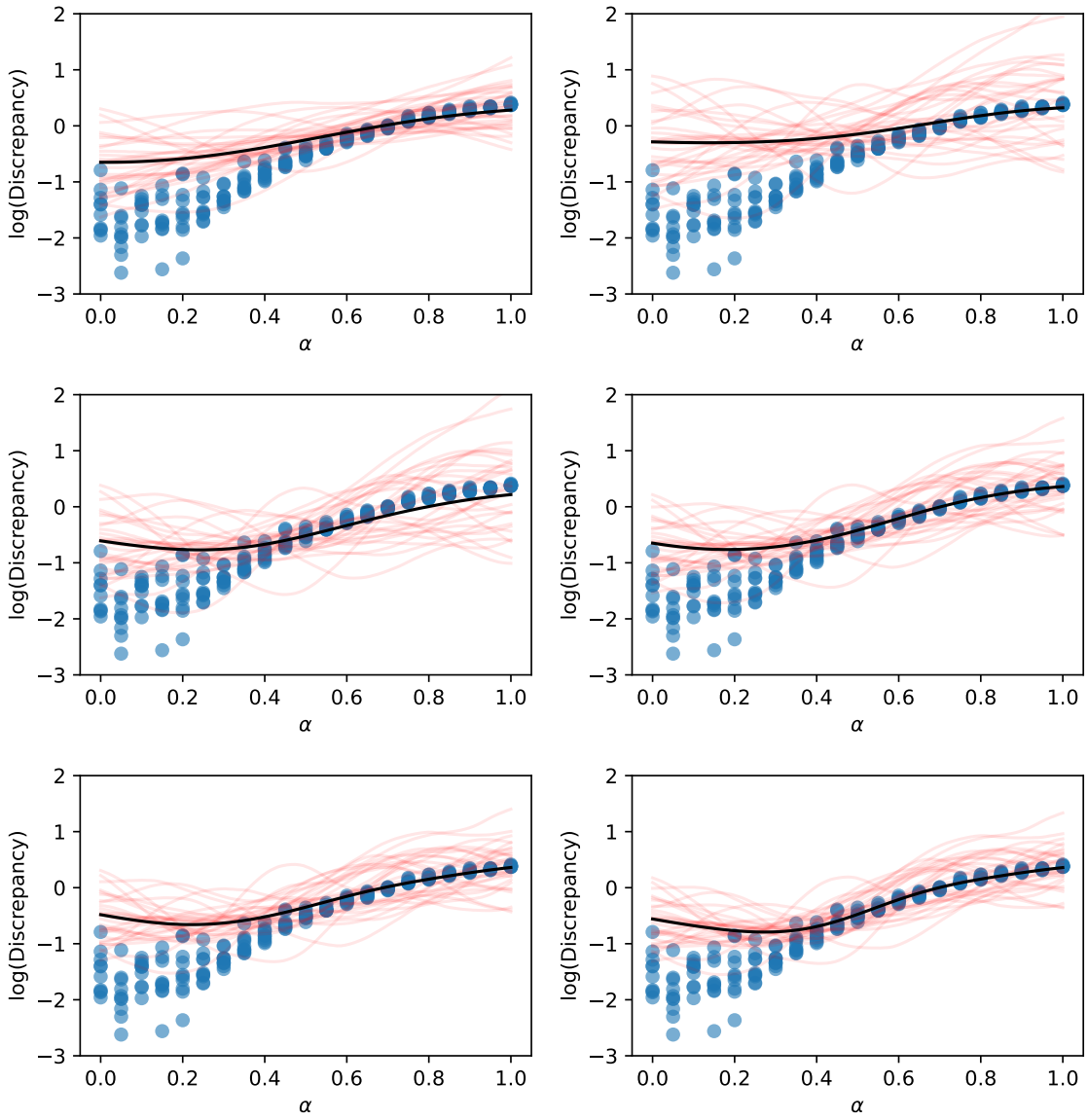


Figure 8.1: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only α was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.

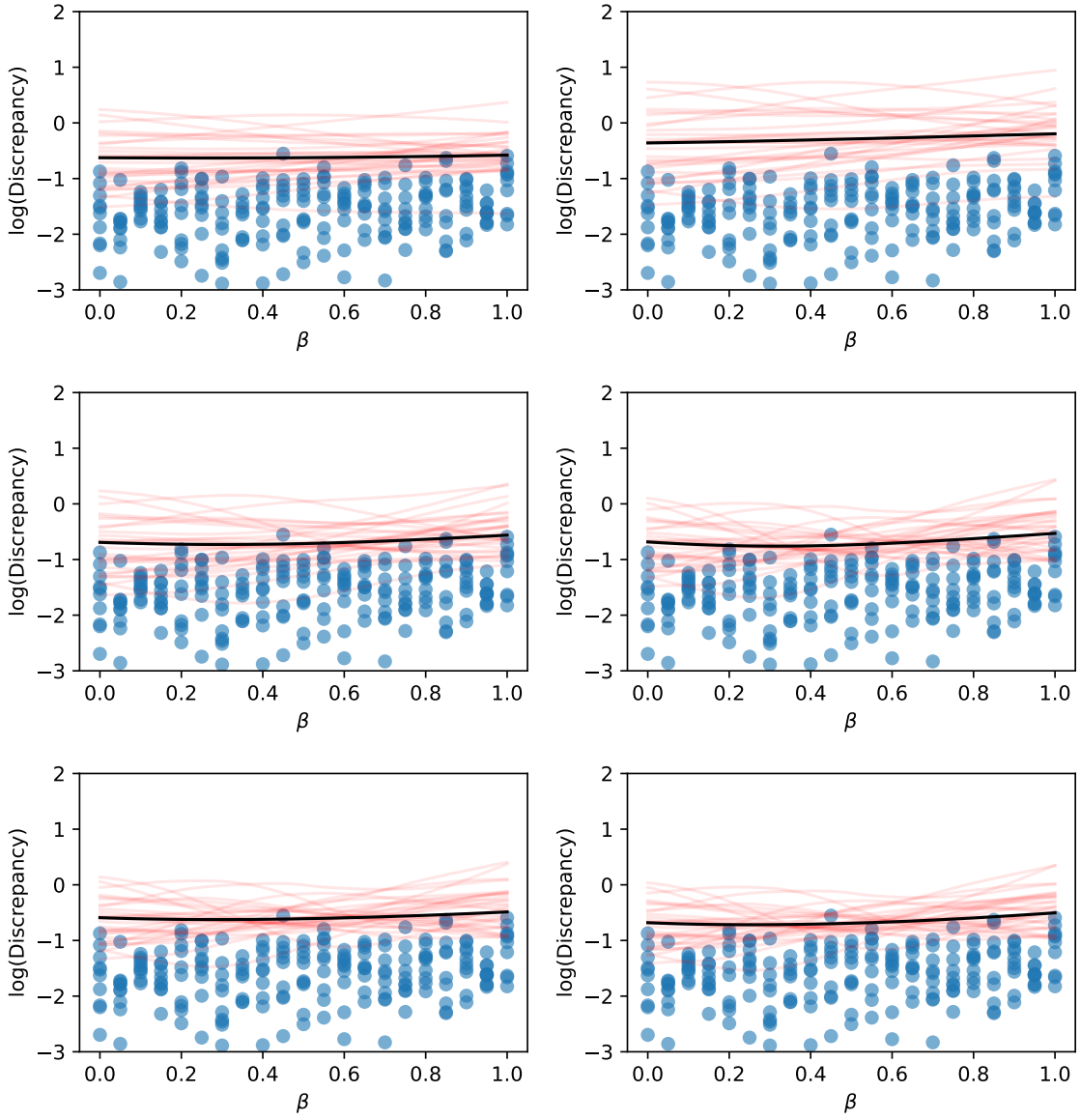


Figure 8.2: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only β was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.

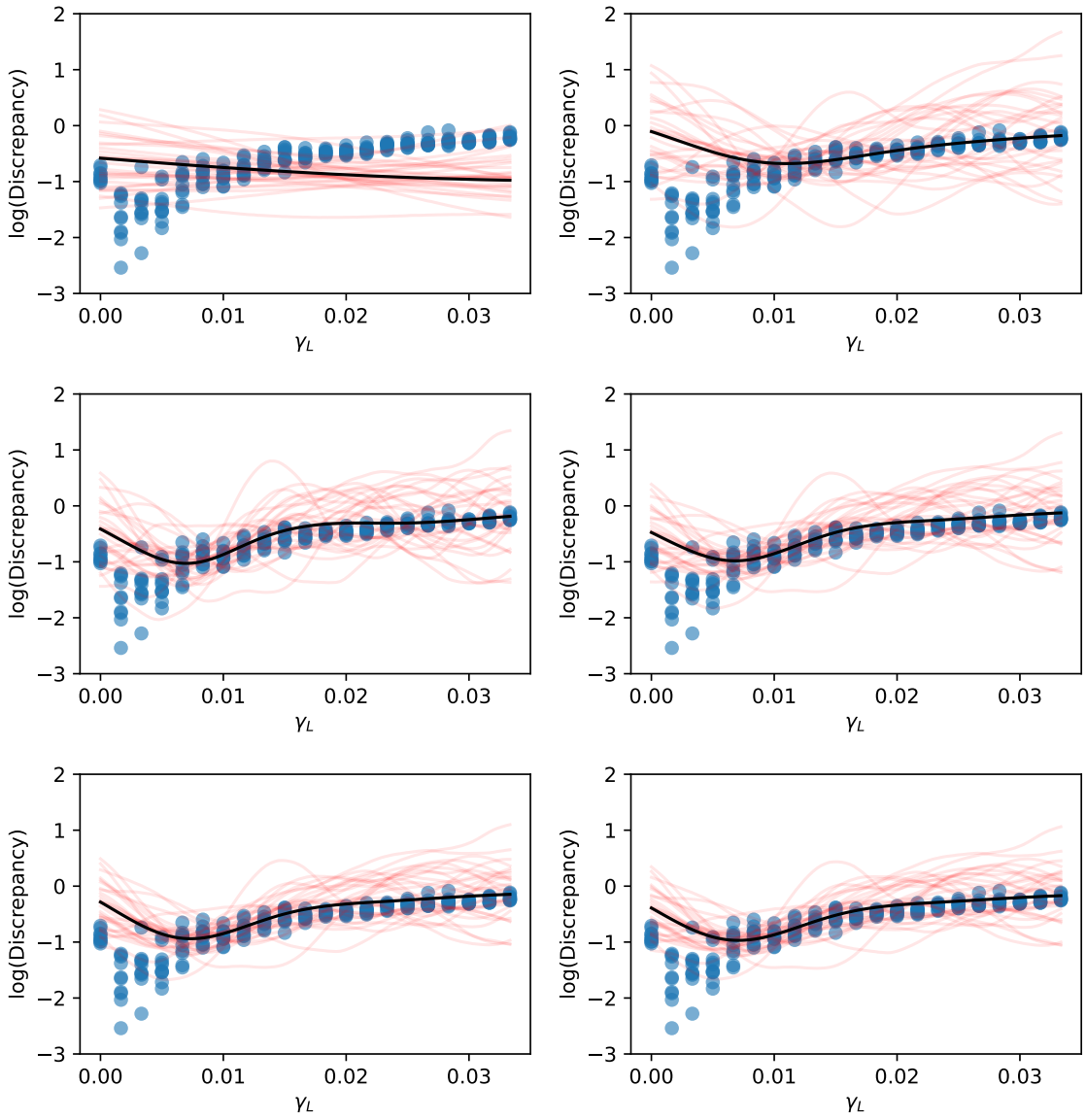


Figure 8.3: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only γ_L was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.

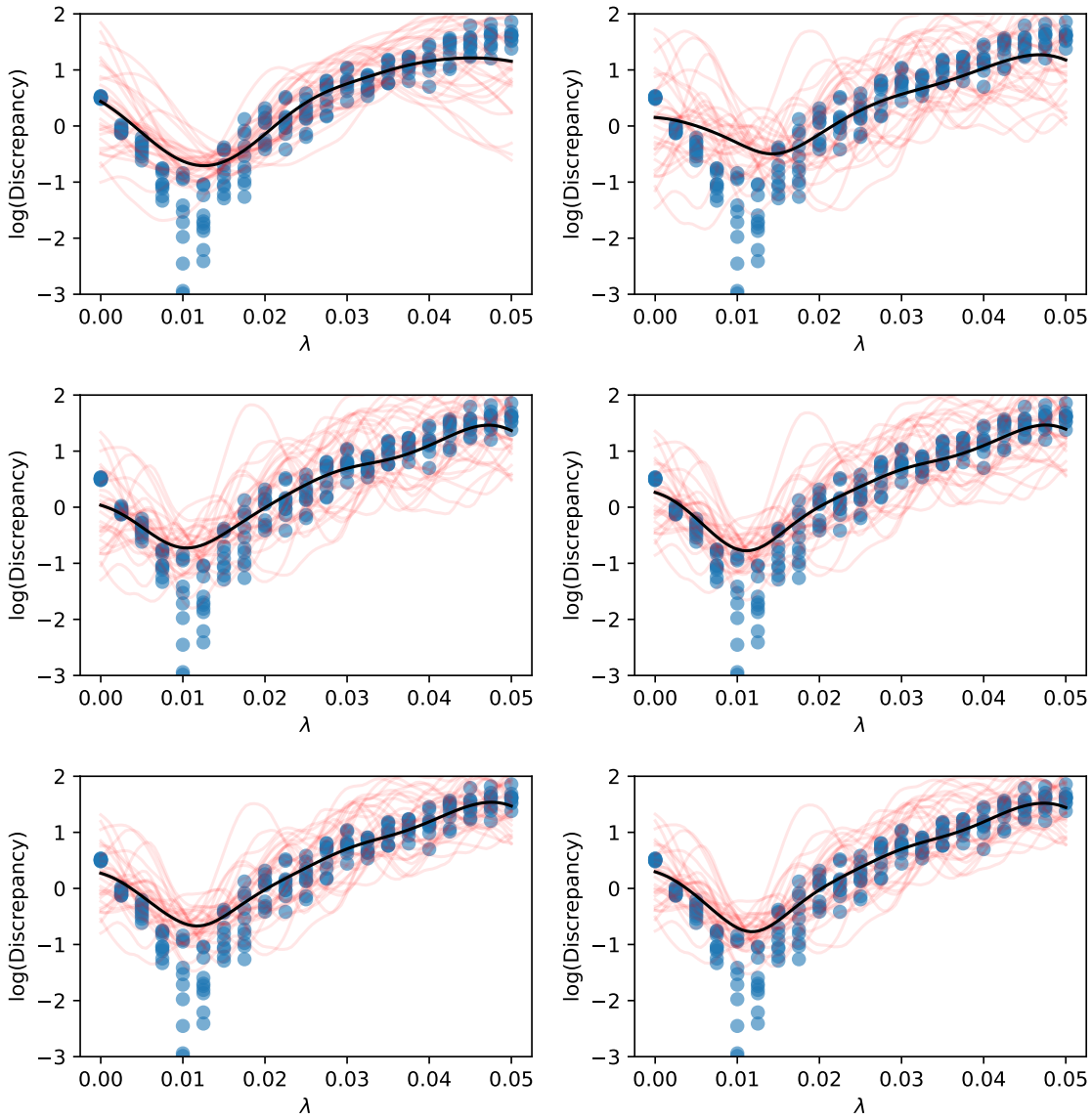


Figure 8.4: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only λ was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.

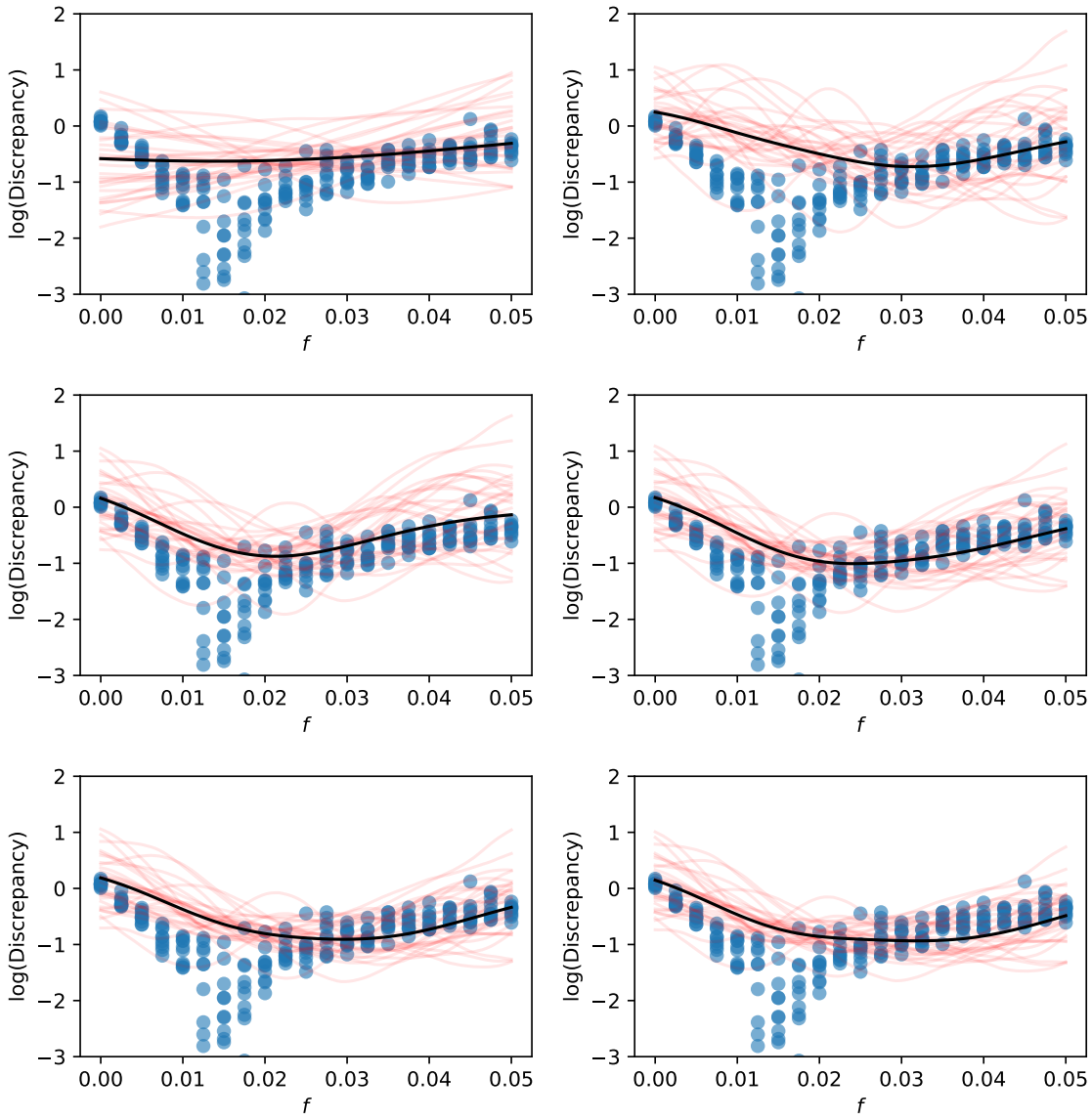


Figure 8.5: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only f was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.

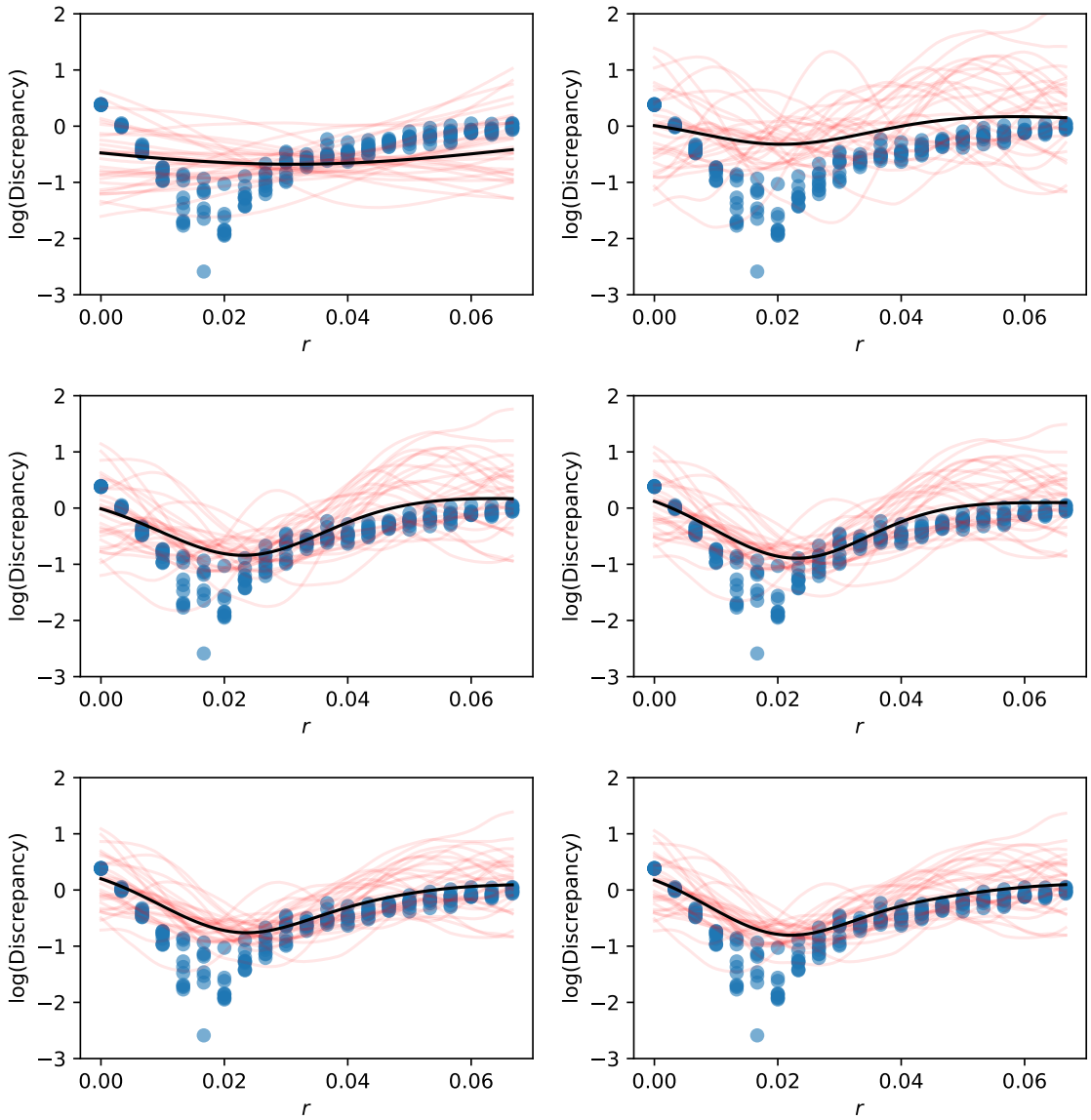


Figure 8.6: $d^{(t)}_{\mathcal{GP}}(\boldsymbol{\theta})$ approximation of $\mathbb{E}(\ln \mathcal{D}(\boldsymbol{\theta}))$, for $t = 0$, 100, 200, 300, 400, and 500. Only r was varied. All other parameters were fixed at the true values. Black line is $\mathbb{E}(d^{(i)}(\boldsymbol{\theta}))$. Blue dots are realisations from $\ln \mathcal{D}(\boldsymbol{\theta})$.