

Efficient Approximation of the Likelihood for Complex Models

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

June 2024



P. vivax Malaria

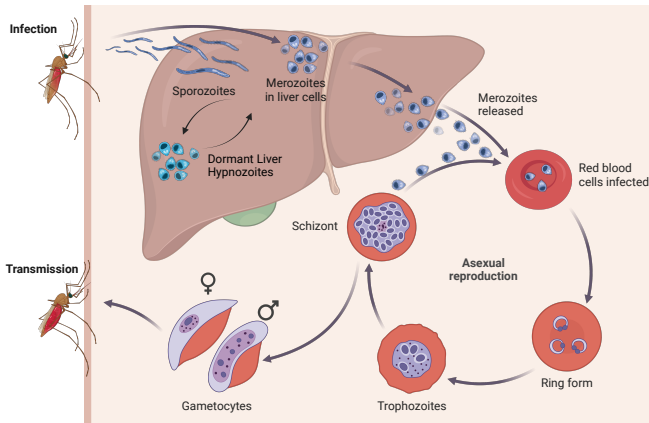


Figure: *P. vivax* life cycle. Dormant liver stage leads to relapses. Created with BioRender.com

P. vivax Malaria

- ▶ Dormant liver stage = more parameters

Vivax Model - Champagne et. al

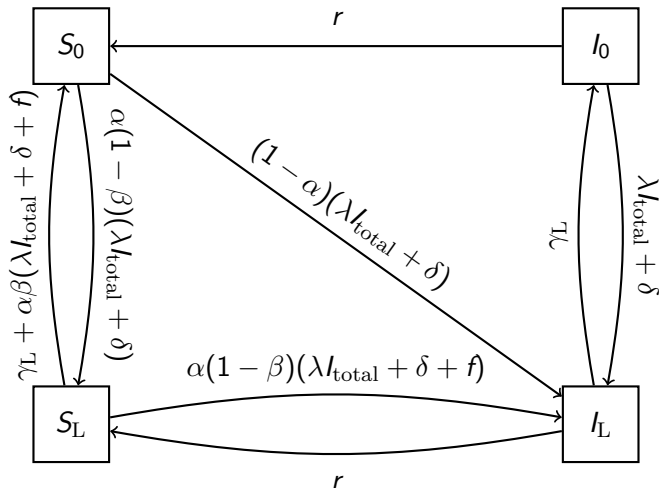


Figure: Champagne et al. 2022 *P. vivax* model

Champagne Model Parameters

- ▶ α : proportion of those infected but cleared of blood stage infections (through treatment)
- ▶ β : a further proportion that are also cleared of liver stage parasites, given that they were also cleared of blood stage infection (radical cure)
- ▶ λ : the rate of infection
- ▶ γ_L : rate of clearance of liver stage disease
- ▶ f : rate of relapse
- ▶ r : rate of blood stage clearance
- ▶ $\delta = 0$ importation rate (fixed)

Ordinary Differential Equations - Champagne et. al

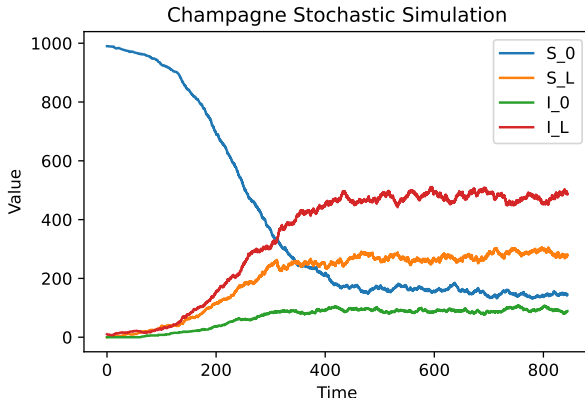
$$\begin{aligned}\frac{dI_L}{dt} = & (1 - \alpha)(\lambda I_{\text{total}} + \delta)(S_0 + S_L) + (\lambda I_{\text{total}} + \delta)I_0 \\ & + (1 - \alpha)fS_L - \gamma_L I_L - rI_L\end{aligned}$$

$$\frac{dI_0}{dt} = -(\lambda I_{\text{total}} + \delta)I_0 + \gamma_L I_L - rI_0$$

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

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

Example Simulation



The Problem

- ▶ How to parametrise such a model.

The Problem

- ▶ How to parametrise such a model.
- ▶ Simulations take long time (and models get a lot more complicated)

Notation

- ▶ θ vector of parameters - e.g. $[\alpha, \beta, \gamma_L, \lambda, f, r]^T$
- ▶ \mathbf{Y}_{obs} : a (summary) vector of observed data e.g. (weekly) incidence, prevalence, (monthly) hospitalisations

Notation

- ▶ θ vector of parameters - e.g. $[\alpha, \beta, \gamma_L, \lambda, f, r]^T$
- ▶ \mathbf{Y}_{obs} : a (summary) vector of observed data e.g. (weekly) incidence, prevalence, (monthly) hospitalisations
- ▶ \mathbf{Y}_{θ} : a random vector of model statistics for given θ .

In an ideal world...

- ▶ There would be an explicit form for the likelihood:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\boldsymbol{\theta})$$

In an ideal world...

- ▶ There would be an explicit form for the likelihood:

$$\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\theta)$$

- ▶ $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|\mathbf{Y}_{\text{obs}})$

In an ideal world...

- ▶ There would be an explicit form for the likelihood:

$$\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\theta)$$

- ▶ $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|\mathbf{Y}_{\text{obs}})$
- ▶ $\Pr(\theta|\mathbf{Y}_{\text{obs}}) \propto \Pr(\mathbf{Y}_{\text{obs}}|\theta) \Pr(\theta)$

In an ideal world...

- ▶ There would be an explicit form for the likelihood:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\boldsymbol{\theta})$$

- ▶ $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$
- ▶ $\Pr(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) \propto \Pr(\mathbf{Y}_{\text{obs}}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$
- ▶ Off to the pub

Or not...

- ▶ Explicit likelihoods often don't exist/are intractible
 - ▶ Champagne model
 - ▶ Agent based models.

A Standard Bayesian Solution

- ▶ Approximate Bayesian Computation (ABC)
 1. Sample θ_i from prior
 2. Run model and observe \mathbf{Y}_{θ_i}
 3. Accept or reject θ_i run based on how well \mathbf{Y}_{θ_i} 'matches' \mathbf{Y}_{obs} .

What is 'matches'

1. $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$

What is 'matches'

1. $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$
 - ▶ Good luck...

What is 'matches'

1. $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$
 - ▶ Good luck...
2. Rescale \mathbf{Y} s, and use discrepancy function $D : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$
e.g. p -norm

$$D(\mathbf{Y}_{\theta_i}, \mathbf{Y}_{\text{obs}}) := \left(\sum_{j=1}^d |\{\mathbf{Y}_{\theta_i}\}_j - \{\mathbf{Y}_{\text{obs}}\}_j|^p \right)^{1/p}$$

Discrepancy Function

$\mathcal{D}(\boldsymbol{\theta}) := D(\mathbf{Y}_{\boldsymbol{\theta}}, \mathbf{Y}_{\text{obs}})$ how 'close' our model is to the observed data using parameters $\boldsymbol{\theta}$

1. Sample θ_i from prior
2. Run model
3. Accept θ_i if $\mathcal{D}(\theta_i) < \varepsilon$.

Overall Idea of my Research

- ▶ ABC fixes one problem but leaves another:
 - ▶ Don't need $\mathcal{L}(\theta)$.
 - ▶ Evaluating $\mathcal{D}(\theta)$ takes as long as a model run.

Overall Idea of my Research

- ▶ ABC fixes one problem but leaves another:
 - ▶ Don't need $\mathcal{L}(\theta)$.
 - ▶ Evaluating $\mathcal{D}(\theta)$ takes as long as a model run.
- ▶ $\mathcal{D}(\theta), \mathcal{D}(\theta')$ will be highly correlated when θ is near θ' .
 - ▶ Gaussian Processes

Gaussian Process Setup

A common assumption is that

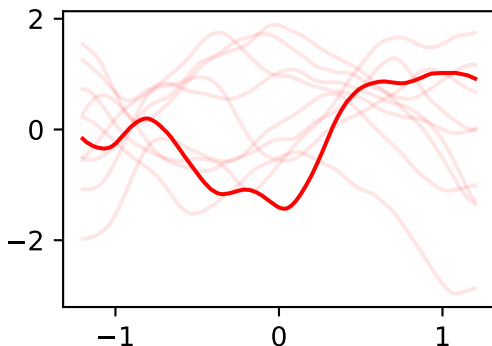
$$\begin{bmatrix} \mathbb{E}(\mathcal{D}(\boldsymbol{\theta}_1)) \\ \vdots \\ \mathbb{E}(\mathcal{D}(\boldsymbol{\theta}_n)) \end{bmatrix} \sim \text{MVN}(\mathbf{0}, \mathbf{K})$$

where $\mathbf{K}_{ij} = k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ for some covariance kernel k that decays to 0 as $\boldsymbol{\theta}_i$ is further away than $\boldsymbol{\theta}_j$.

Gaussian Processes on \mathbb{R}^d

Definition (Gaussian Process)

A collection of random variables $\{f(\mathbf{x})\}_{\mathbf{x} \in \mathbb{R}^d}$ is a Gaussian process if all finite dimensional distributions are multivariate normal distributed.



Equivalent GP definition

Definition (Gaussian Process)

A collection of random variables $\{f(\mathbf{x})\}_{\mathbf{x} \in \mathbb{R}^d}$ is a Gaussian process if there is a function $m : \mathcal{X} \rightarrow \mathbb{R}$ and covariance kernel $k : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ such that for all $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} \right)$$

where $\mathbf{K}_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$.

k Determines Smoothness

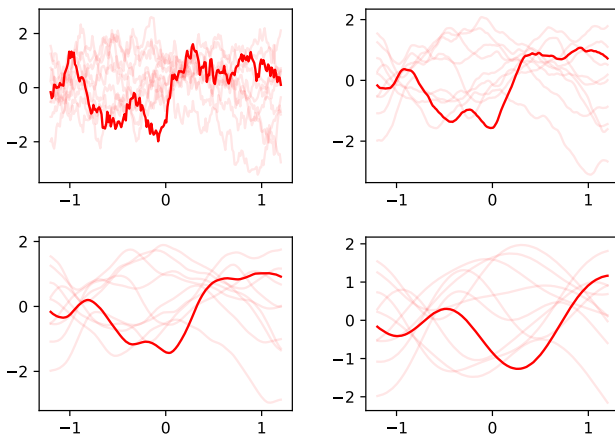


Figure: Matérn 1/2, 3/2, 5/2, and squared exponential kernels.

Gaussian Process Continuity

- ▶ Induce continuity by forcing $k(\mathbf{x}, \mathbf{x}') \rightarrow \text{Var}(f(\mathbf{x}))$ (hence $\text{Cor}(f(\mathbf{x}), f(\mathbf{x}')) \rightarrow 1$) as $\mathbf{x} \rightarrow \mathbf{x}'$.

Common Covariance Kernels

- ▶ Matérn Kernel with hyperparameter $\nu : \lfloor \nu \rfloor$ times mean square differentiable.
- ▶ $\nu \rightarrow \infty$ - infinitely mean square differentiable squared exponential covariance kernel (strong assumption)

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

k Determines Class of Functions

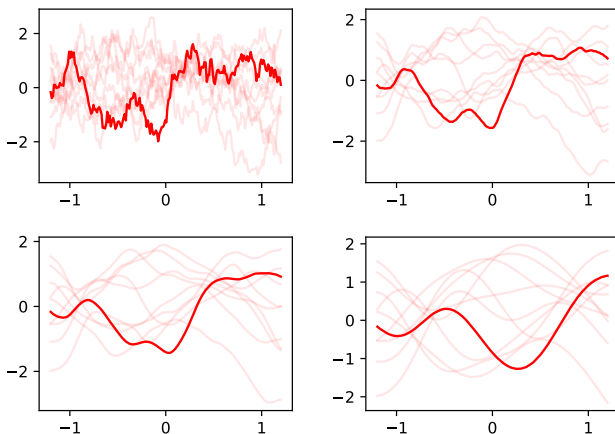
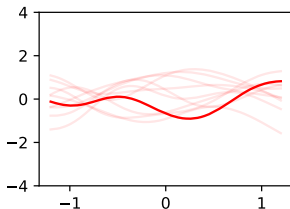


Figure: Matérn 1/2, 3/2, 5/2, and squared exponential kernels.

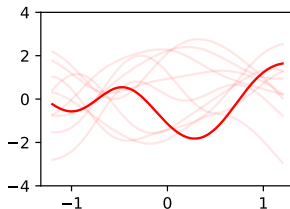
Kernel Hyperparameters

- ▶ Matérn and squared exponential kernel can both be written in the form $k(\mathbf{x}, \mathbf{x}') = \sigma_k^2 \kappa(\|\mathbf{x}, \mathbf{x}'\|/\ell)$
- ▶ $1/\ell$ rate of covariance decay
- ▶ $\sigma_k = \text{Var}(f(\mathbf{x}))$

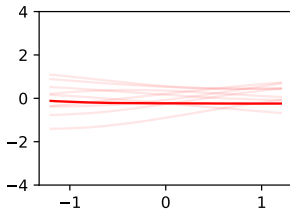
Kernel Hyperparameters



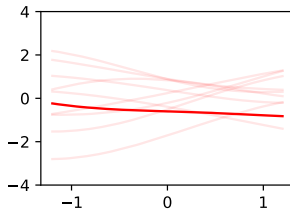
(a) $\ell = 1/2$ $\sigma_k^2 = 1/2$



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



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



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

Discrepancy Function Context

- ▶ Long term play: use a Gaussian process as a surrogate model for $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$

Discrepancy Function Context

- ▶ Long term play: use a Gaussian process as a surrogate model for $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$
- ▶ What if we have observations already?

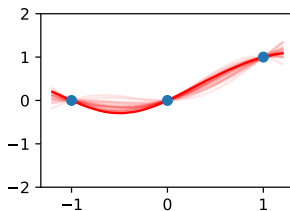
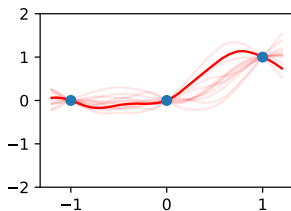
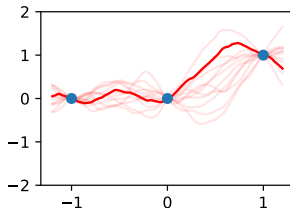
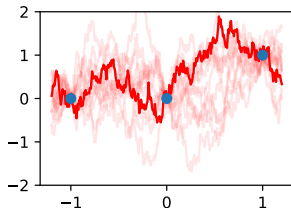
Gaussian Process Regression

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

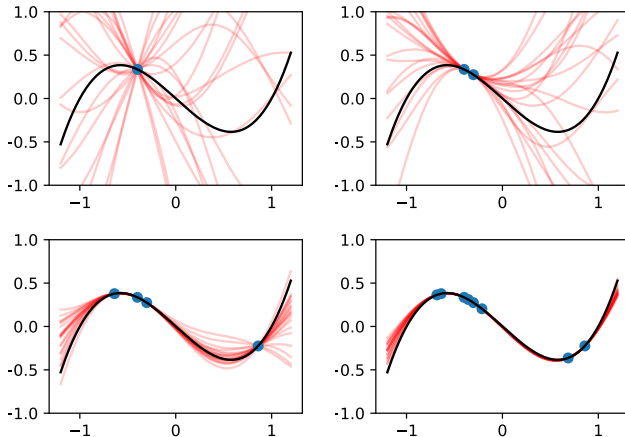
implies

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

Conditioning Gaussian Processes



GP regression on $x(x-1)(x+1)$

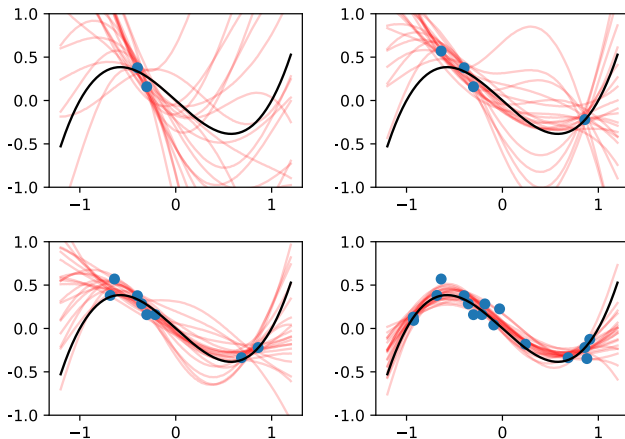


Normal observation noise

If observations actually $f(\mathbf{x}_i) + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma_o^2)$ i.i.d., then

$$\begin{bmatrix} f(\mathbf{x}_1) + \varepsilon_1 \\ \vdots \\ f(\mathbf{x}_n) + \varepsilon_n \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} + \sigma_o^2 \mathbf{I}_n \right)$$

GP regression on $x(x-1)(x+1) + \varepsilon$, $\varepsilon \sim N(0, \sigma_o^2)$



Key Assumptions

- ▶ $\mathcal{D}(\boldsymbol{\theta}) \stackrel{d}{\approx} \mathcal{D}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}, \boldsymbol{\theta}'$ close.
- ▶ $\mathcal{D}(\boldsymbol{\theta})$ approximately distributed $N(\cdot, \sigma_0^2)$ with σ_0^2 independent of $\boldsymbol{\theta}$.
- ▶ $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$ can be well approximated by a Gaussian process.

Key Idea

Approximate $\mathcal{D}(\boldsymbol{\theta})$ with $\mathcal{D}_{GP}(\boldsymbol{\theta})$, a Gaussian process with observation noise.

1. Sample θ_i from prior
2. Run model
3. Accept θ_i if $\mathcal{D}(\theta_i) < \varepsilon$.

Approximate ABC...??

1. Sample θ_i from prior
2. Run model
3. Accept θ_i if $\mathcal{D}_{\mathcal{GP}}(\theta_i) < \varepsilon$.

Synthetic Likelihood

The probability of drawing and accepting θ under the ABC is

$$\Pr(\mathcal{D}_{\mathcal{GP}}(\theta) < \varepsilon) \Pr(\theta)$$

and hence we consider $L(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathcal{D}_{\mathcal{GP}}(\theta) < \varepsilon)$ a synthetic likelihood - an approximation of the true likelihood $\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}})$

Log Gaussian Process

- ▶ Alternatively we can model $\ln \mathcal{D}(\boldsymbol{\theta})$ as a Gaussian process $d_{\mathcal{GP}}(\boldsymbol{\theta})$.

Log Gaussian Process

- ▶ Alternatively we can model $\ln \mathcal{D}(\boldsymbol{\theta})$ as a Gaussian process $d_{\mathcal{GP}}(\boldsymbol{\theta})$.
- ▶ Key assumptions become:
 - ▶ $\ln \mathcal{D}(\boldsymbol{\theta}) \stackrel{d}{\approx} \ln \mathcal{D}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}, \boldsymbol{\theta}'$ close.
 - ▶ $\mathcal{D}(\boldsymbol{\theta})$ approximately distributed $LN(\cdot, \sigma_0^2)$ with σ_0^2 independent of $\boldsymbol{\theta}$.
 - ▶ $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})]$ can be well approximated by a Gaussian process.

Log Gaussian Process

- ▶ Alternatively we can model $\ln \mathcal{D}(\boldsymbol{\theta})$ as a Gaussian process $d_{\mathcal{GP}}(\boldsymbol{\theta})$.
- ▶ Key assumptions become:
 - ▶ $\ln \mathcal{D}(\boldsymbol{\theta}) \stackrel{d}{\approx} \ln \mathcal{D}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}, \boldsymbol{\theta}'$ close.
 - ▶ $\mathcal{D}(\boldsymbol{\theta})$ approximately distributed $LN(\cdot, \sigma_0^2)$ with σ_0^2 independent of $\boldsymbol{\theta}$.
 - ▶ $\mathbb{E}[\ln \mathcal{D}(\boldsymbol{\theta})]$ can be well approximated by a Gaussian process.
- ▶

$$\Pr(\mathcal{D}(\boldsymbol{\theta}_i) < \varepsilon) \approx \Pr(d_{\mathcal{GP}}(\boldsymbol{\theta}_i) < \ln \varepsilon)$$

Where to sample $\mathcal{D}(\theta)$

- ▶ To generate a reliable $\mathcal{D}_{\mathcal{GP}}$, we need to sample widely
- ▶ Generating $\mathcal{D}(\theta)$ still costly...
- ▶ Therefore sample where:

Where to sample $\mathcal{D}(\theta)$

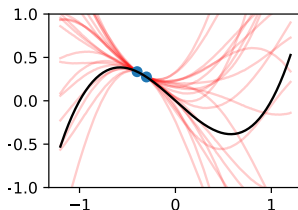
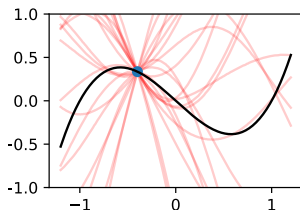
- ▶ To generate a reliable $\mathcal{D}_{\mathcal{GP}}$, we need to sample widely
- ▶ Generating $\mathcal{D}(\theta)$ still costly...
- ▶ Therefore sample where:
 - ▶ $\mathcal{D}(\theta)$ small.

Where to sample $\mathcal{D}(\theta)$

- ▶ To generate a reliable $\mathcal{D}_{\mathcal{GP}}$, we need to sample widely
- ▶ Generating $\mathcal{D}(\theta)$ still costly...
- ▶ Therefore sample where:
 - ▶ $\mathcal{D}(\theta)$ small.
 - ▶ $\mathcal{D}(\theta)$ highly unknown.

Where to sample $\mathcal{D}(\theta)$

- ▶ To generate a reliable \mathcal{D}_{GP} , we need to sample widely
- ▶ Generating $\mathcal{D}(\theta)$ still costly...
- ▶ Therefore sample where:
 - ▶ $\mathcal{D}(\theta)$ small.
 - ▶ $\mathcal{D}(\theta)$ highly unknown.



Bayesian Acquisition Functions

- Quantify where we expect $\mathcal{D}(\theta)$ small and highly unknown with Bayesian acquisition function A , and choose $\arg \min_{\theta} A(\theta)$

Bayesian Acquisition Functions

- ▶ Gutmann and Cor 2016 use lower confidence bound

$$A_{\text{LCB}}(\boldsymbol{\theta}) := \mu(\boldsymbol{\theta}) - \eta_t \sqrt{v(\boldsymbol{\theta})}$$

- ▶ $\mu(\boldsymbol{\theta})$, $v(\boldsymbol{\theta})$ are posterior mean and variance of $D_{\mathcal{GP}}(\boldsymbol{\theta})$
- ▶ η_t a slowly increasing function in t , the number of previous samples

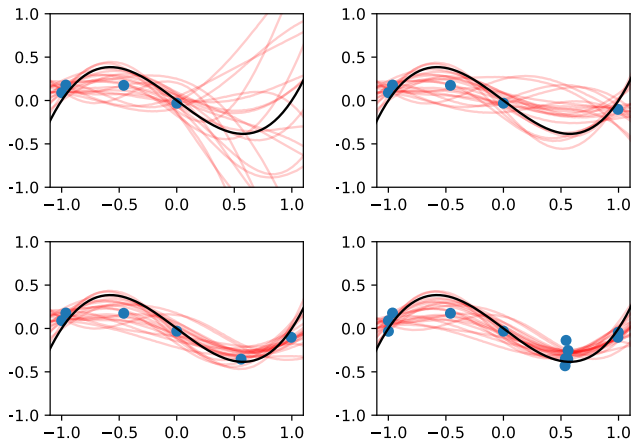
Bayesian Acquisition Functions

- ▶ Expected information

$$A_{\text{EI}}(\boldsymbol{\theta}) := \mathbb{E}[\min(\mu_{\min} - \mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta}), 0)]$$

- ▶ $\mu_{\min} := \min_{\boldsymbol{\theta}} \mu(\boldsymbol{\theta})$

Lower Confidence Bound



Vivax Model - Champagne et. al

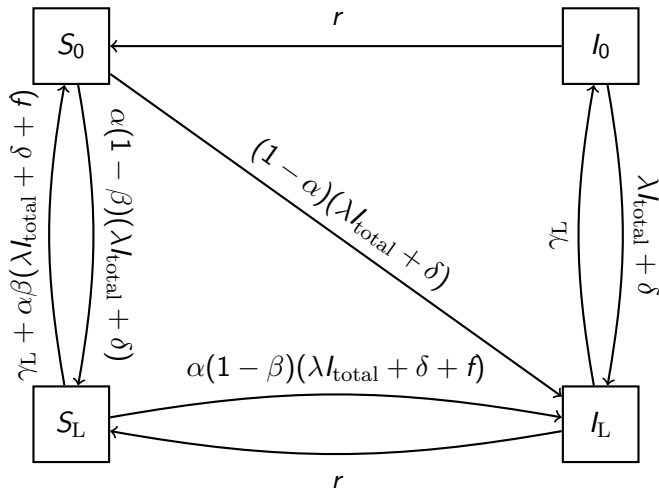


Figure: Champagne et al. 2022 *P. vivax* model

Champagne Model Parameters

- ▶ α : proportion of those infected but cleared of blood stage infections (through treatment)
- ▶ β : a further proportion that are also cleared of liver stage parasites, given that they were also cleared of blood stage infection (radical cure)
- ▶ λ : the rate of infection
- ▶ γ_L : rate of clearance of liver stage disease
- ▶ f : rate of relapse
- ▶ r : rate of blood stage clearance

'Observed' Data

- ▶ 'Observed' data from one simulation of 10 initial infections, in a population of 1000 people using the parameters reported in Champagne et al. 2022.
- ▶ $\mathbf{Y}_{\text{obs}} := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$
 - ▶ w_{obs} : weekly incidence around (stochastic) equilibrium
 - ▶ p_{obs} : prevalence around (stochastic) equilibrium
 - ▶ m_{obs} : incidence in the first month of the epidemic

'Observed' Data

- ▶ 'Observed' data from one simulation of 10 initial infections, in a population of 1000 people using the parameters reported in Champagne et al. 2022.
- ▶ $\mathbf{Y}_{\text{obs}} := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$
 - ▶ w_{obs} : weekly incidence around (stochastic) equilibrium
 - ▶ p_{obs} : prevalence around (stochastic) equilibrium
 - ▶ m_{obs} : incidence in the first month of the epidemic

'Observed' Data

- ▶ 'Observed' data from one simulation of 10 initial infections, in a population of 1000 people using the parameters reported in Champagne et al. 2022.
- ▶ $\mathbf{Y}_{\text{obs}} := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$
 - ▶ w_{obs} : weekly incidence around (stochastic) equilibrium
 - ▶ p_{obs} : prevalence around (stochastic) equilibrium
 - ▶ m_{obs} : incidence in the first month of the epidemic
- ▶ $\mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r)$ is the L_2 norm of the relative differences

$$\sqrt{\left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2 + \left(\frac{m - m_{\text{obs}}}{m_{\text{obs}}}\right)^2 + \left(\frac{w - w_{\text{obs}}}{w_{\text{obs}}}\right)^2}$$

GP choices

- ▶ $\mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r)$ is the L_2 norm of the relative differences

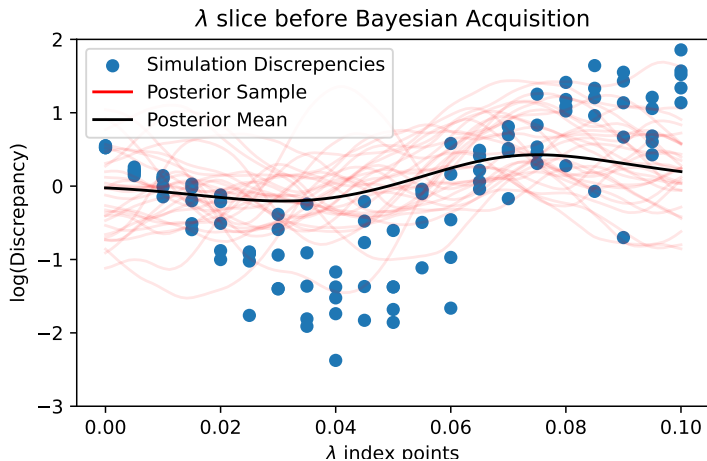
$$\sqrt{\left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2 + \left(\frac{m - m_{\text{obs}}}{m_{\text{obs}}}\right)^2 + \left(\frac{w - w_{\text{obs}}}{w_{\text{obs}}}\right)^2}$$

- ▶ \mathcal{GP} choices
 - ▶ Modelled In \mathcal{D} as a Gaussian process
 - ▶ Matern kernel with $\nu = 5/2$
 - ▶ $\ell, \sigma_k^2, \sigma_o^2$ selected by leave one out cross validation.

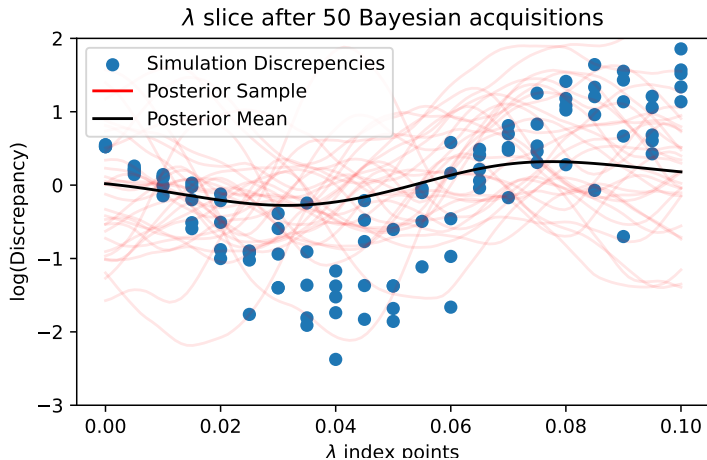
How did it go?



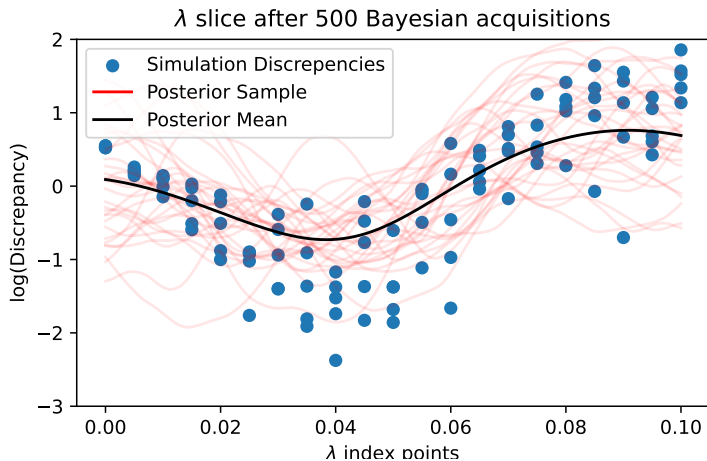
How did it go?



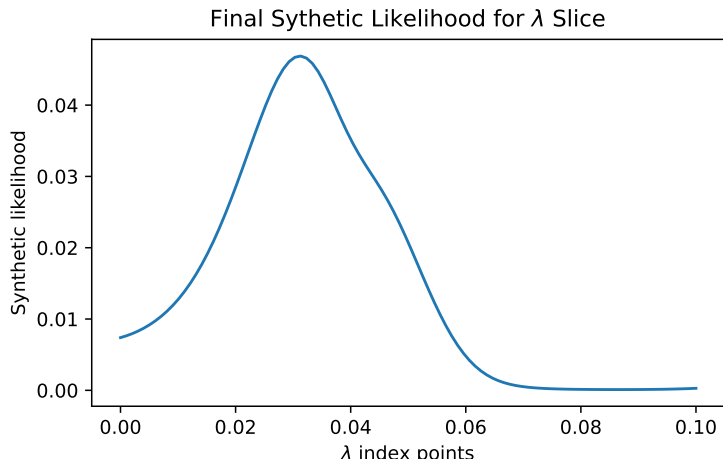
How did it go?



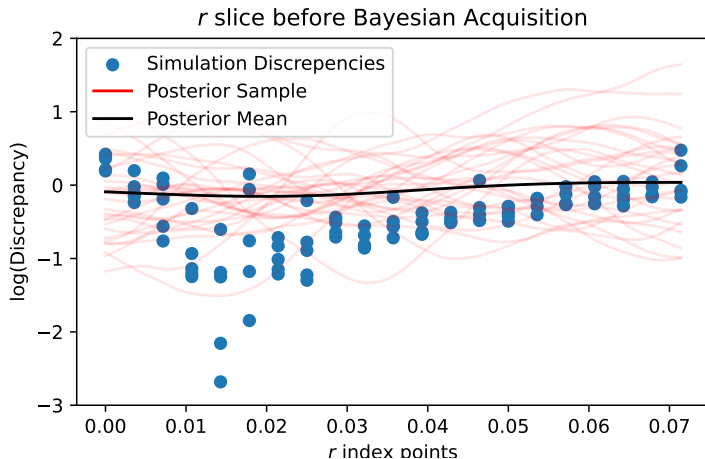
How did it go?



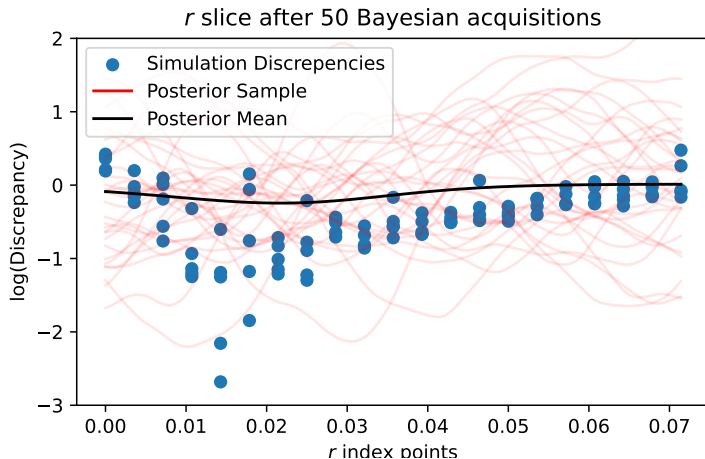
How did it go?



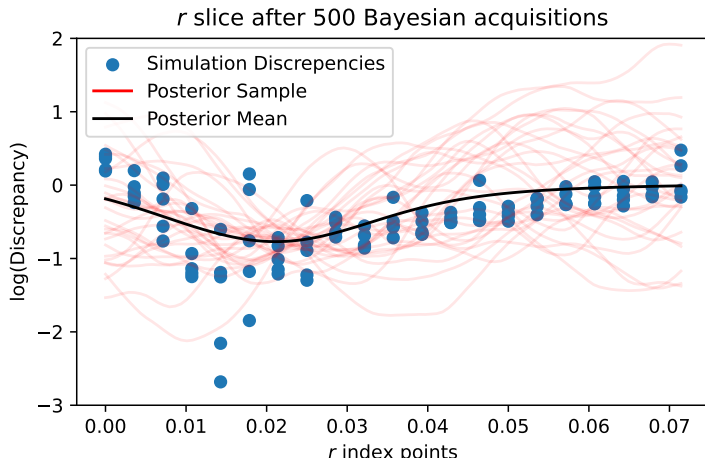
How did it go?



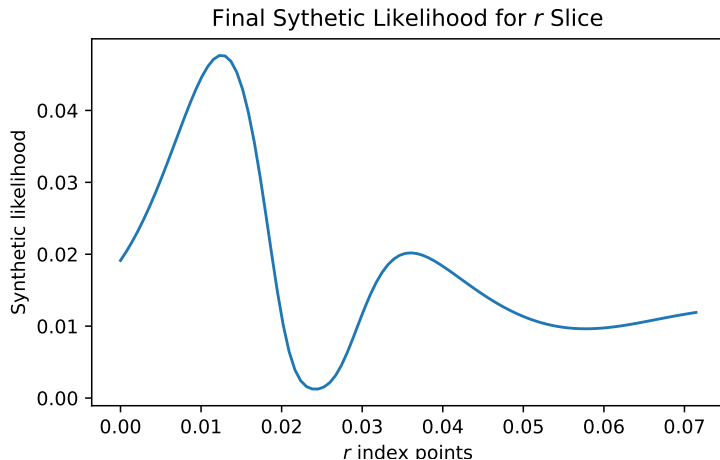
How did it go?



How did it go?



How did it go?



Discussion

- ▶ Observation variance is considered constant across the GP (or log GP)
 - ▶ Particularly a problem at the threshold
- ▶ Assumes that normal/log-normal distribution approximates $\mathcal{D}(\theta)$
- ▶ Jumps where there is threshold/bifurcation behaviour
 - ▶ Student t -Process?

Thanks to

- ▶ Eamon Conway
- ▶ Jennifer Flegg



Bibliography



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).



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).

