

# Efficient Approximation of the Likelihood for Complex Models

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

May 2024



# Notation

- ▶ Model a (random) function  $f: \Theta \rightarrow \mathcal{Y}$ .
  - ▶  $\Theta$  : parameter space
  - ▶  $\mathcal{Y}$  : model output space
  - ▶  $\mathbf{Y}_\theta := f(\theta)$  (assumed same form as  $\mathbf{Y}_{\text{obs}}$ ).

# Notation

- ▶ Model a (random) function  $f: \Theta \rightarrow \mathcal{Y}$ .
  - ▶  $\Theta$  : parameter space
  - ▶  $\mathcal{Y}$  : model output space
  - ▶  $\mathbf{Y}_\theta := f(\theta)$  (assumed same form as  $\mathbf{Y}_{\text{obs}}$ ).
- ▶  $\mathbf{Y}_{\text{obs}}$  : a vector of observed data (incidence, prevalence, hospitalisations etc.)

# Notation

- ▶ Model a (random) function  $f: \Theta \rightarrow \mathcal{Y}$ .
  - ▶  $\Theta$  : parameter space
  - ▶  $\mathcal{Y}$  : model output space
  - ▶  $\mathbf{Y}_\theta := f(\theta)$  (assumed same form as  $\mathbf{Y}_{\text{obs}}$ ).
- ▶  $\mathbf{Y}_{\text{obs}}$  : a vector of observed data (incidence, prevalence, hospitalisations etc.)
- ▶  $S(\mathbf{Y}_{\text{obs}})$  : summary statistic (vector) of observed data (average weekly incidence etc.)

# Parameter inference would become easy if we had...

- ▶ An explicit form for the likelihood:
  - ▶  $\mathcal{L}(\theta | \mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}} | \theta)$

# Parameter inference would become easy if we had...

- ▶ An explicit form for the likelihood:
  - ▶  $\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\theta)$
  - ▶ Or  $\mathcal{L}(\theta|S(\mathbf{Y}_{\text{obs}})) := \Pr(S(\mathbf{Y}_{\text{obs}})|\theta)$

# Parameter inference would become easy if we had...

- ▶ An explicit form for the likelihood:
  - ▶  $\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\theta)$
  - ▶ Or  $\mathcal{L}(\theta|S(\mathbf{Y}_{\text{obs}})) := \Pr(S(\mathbf{Y}_{\text{obs}})|\theta)$
- ▶  $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|S(\mathbf{Y}_{\text{obs}}))$

# Parameter inference would become easy if we had...

- ▶ An explicit form for the likelihood:
  - ▶  $\mathcal{L}(\theta|\mathbf{Y}_{\text{obs}}) := \Pr(\mathbf{Y}_{\text{obs}}|\theta)$
  - ▶ Or  $\mathcal{L}(\theta|S(\mathbf{Y}_{\text{obs}})) := \Pr(S(\mathbf{Y}_{\text{obs}})|\theta)$
- ▶  $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta|S(\mathbf{Y}_{\text{obs}}))$
- ▶  $\Pr(\theta|S(\mathbf{Y}_{\text{obs}})) \propto \Pr(S(\mathbf{Y}_{\text{obs}})|\theta) \Pr(\theta)$



- ▶ Explicit likelihoods often don't exist/are intractible
  - ▶ eg. agent based models

# A Standard Bayesian Solution

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

# What is 'matches'

- ▶ Option 1:  $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$  or  $S(\mathbf{Y}_{\theta_i}) = S(\mathbf{Y}_{\text{obs}})$

# What is 'matches'

- ▶ Option 1:  $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$  or  $S(\mathbf{Y}_{\theta_i}) = S(\mathbf{Y}_{\text{obs}})$
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ e.g.  $p$ -norm

$$\|S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})\|_p := \left( \sum_{i=1}^d |S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})|^p \right)^{1/p}$$

# What is 'matches'

- ▶ Option 1:  $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$  or  $S(\mathbf{Y}_{\theta_i}) = S(\mathbf{Y}_{\text{obs}})$
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ e.g.  $p$ -norm

$$\|S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})\|_p := \left( \sum_{i=1}^d |S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})|^p \right)^{1/p}$$

- ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).

# What is 'matches'

- ▶ Option 1:  $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$  or  $S(\mathbf{Y}_{\theta_i}) = S(\mathbf{Y}_{\text{obs}})$
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ e.g.  $p$ -norm

$$\|S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})\|_p := \left( \sum_{i=1}^d |S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})|^p \right)^{1/p}$$

- ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).
- ▶  $\mathcal{D}(\theta) := D(S(\mathbf{Y}_{\theta}), S(\mathbf{Y}_{\text{obs}}))$

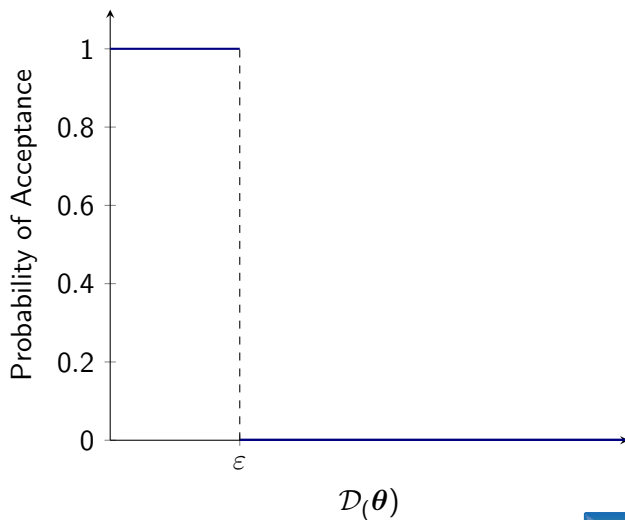
# What is 'matches'

- ▶ Option 1:  $\mathbf{Y}_{\theta_i} = \mathbf{Y}_{\text{obs}}$  or  $S(\mathbf{Y}_{\theta_i}) = S(\mathbf{Y}_{\text{obs}})$
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ e.g.  $p$ -norm

$$\|S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})\|_p := \left( \sum_{i=1}^d |S(\mathbf{Y}_{\theta_i}) - S(\mathbf{Y}_{\text{obs}})|^p \right)^{1/p}$$

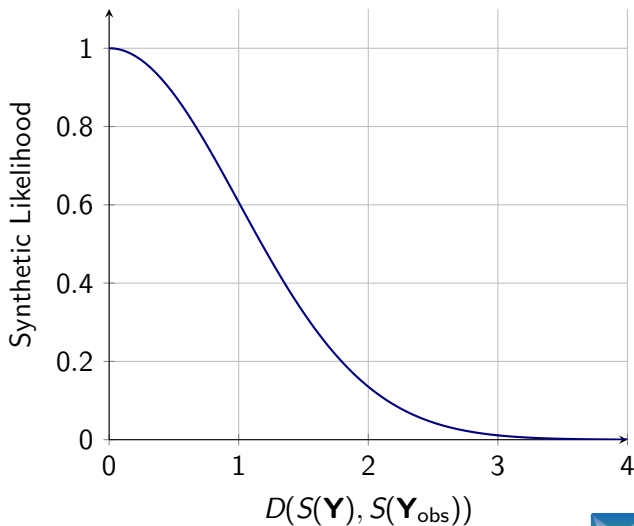
- ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).
- ▶  $\mathcal{D}(\theta) := D(S(\mathbf{Y}_{\theta}), S(\mathbf{Y}_{\text{obs}}))$
- ▶  $\mathcal{D}(\theta_i)$  is 'how close' we were using parameters  $\theta_i$ .

# Uniform Acceptance Probability





# Acceptance Probability



# Overall Idea of my Research

- ▶ Can we predict  $\mathcal{D}(\theta_i)$  without having to evaluate  $f(\theta_i)$ ?

# Overall Idea of my Research

- ▶ Can we predict  $\mathcal{D}(\theta_i)$  without having to evaluate  $f(\theta_i)$ ?
- ▶ Locally, hopefully yes.

# Gaussian Processes

- ▶ Random functions.
- ▶ Common examples - Brownian motion, Ornstein Uhlenbeck process.

# 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. That is, there is a function  $m : \mathcal{X} \rightarrow \mathbb{R}$  and kernel  $k : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  such that for all finite sets  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ,*

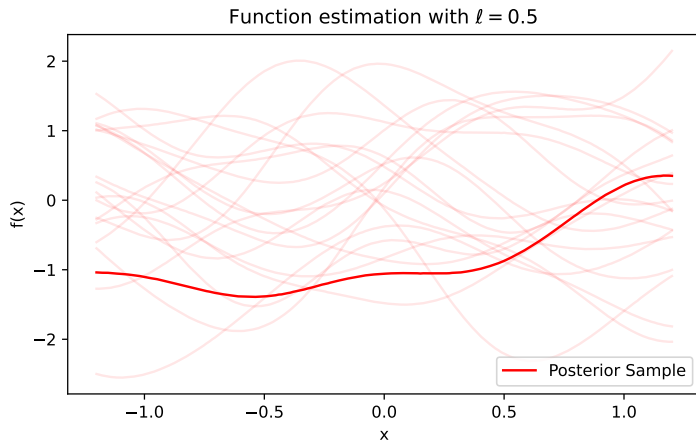
$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} \right)$$

where

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$



# Gaussian Process Example Realisations



# Covariance Kernel Motivation

- ▶ Kernel determines the amount of covariance between sets of indices.
- ▶ When the distance between indices is small, covariance needs to be large

# Common Covariance Kernels

## ▶ Matern Kernel

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

where  $K_{\nu}$  is a modified Bessel function ( $\|\cdot\|$  is the euclidean distance)

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

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



# Kernel Classes

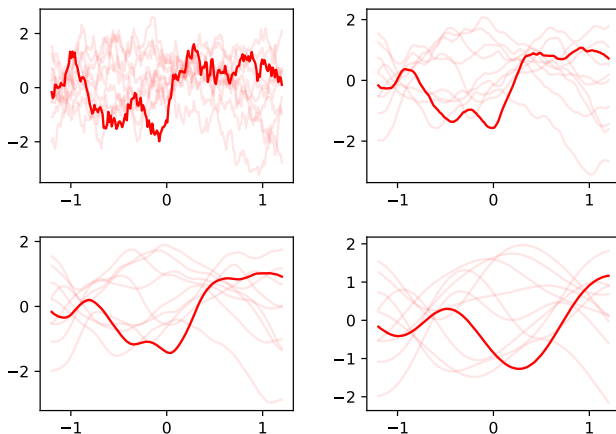


Figure: Matérn 1/2, 3/2, 5/2, and squared exponential kernels.

# Gaussian Process Regression

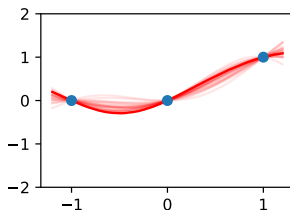
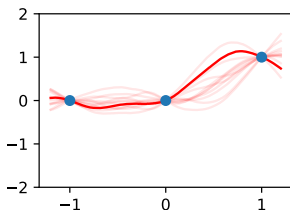
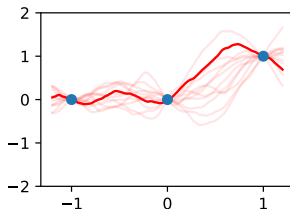
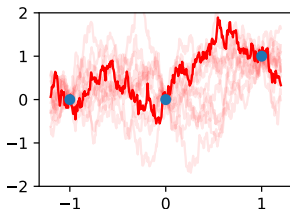
$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \right)$$

implies

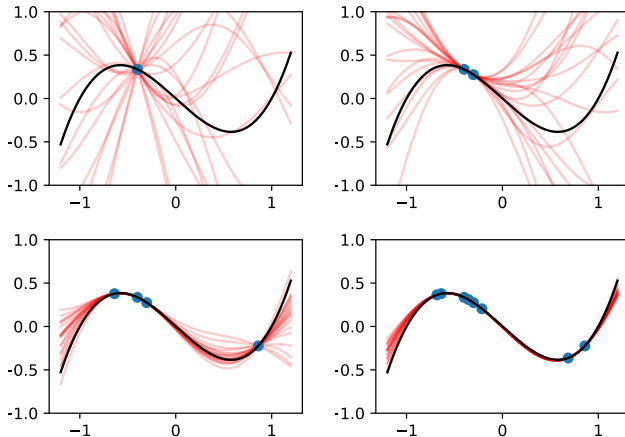
$$f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N} \left( m(\mathbf{x}) + K_* K_{**}^{-1} (f(\mathbf{x}_*) - m(\mathbf{x}_*)), K - K_* K_{**}^{-1} K_*^T \right).$$

# Fitting our GP to data

GPs are 'priors'



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

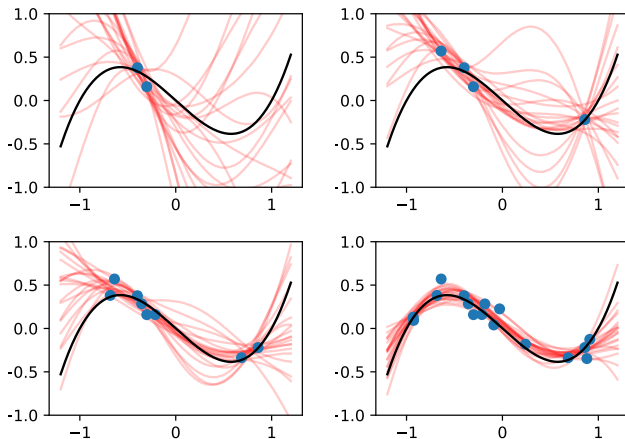


# What if we have noise?

Add observation variance  $\sigma_o^2$ ,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} + \sigma_o^2 \mathbf{I}_n \right)$$

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



# Overall Idea again

- ▶ Can we predict  $\mathcal{D}(\theta_i)$  without having to evaluate  $f(\theta_i)$  <2->
- ▶  $\mathcal{D}(\theta) \approx \mathcal{D}(\theta')$  for  $\theta, \theta'$  close. <3->
- ▶ Approximate  $\mathcal{D}(\theta)$  by a Gaussian process  $\mathcal{D}_{\mathcal{GP}}(\theta)$

# Bayesian Acquisition

- ▶ High expected  $\mathcal{D}_{\mathcal{GP}}(\theta)$  with low variance = waste of time (and resources)
- ▶ Quantify this using a Bayesian acquisition function  $A$ , and choose  $\arg \min_{\theta} A(\theta)$



# Bayesian Acquisition

- ▶ Gutmann and Cor 2016 uses 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
- ▶  $\eta_t := \sqrt{2 \ln[t^{d/2+2} \pi^2 / (3\epsilon)]}$ , with  $\epsilon \in (0, 1)$
- ▶ (Claim of theoretical guarantees = load of rubbish)

# Bayesian Acquisition

## ► Expected information

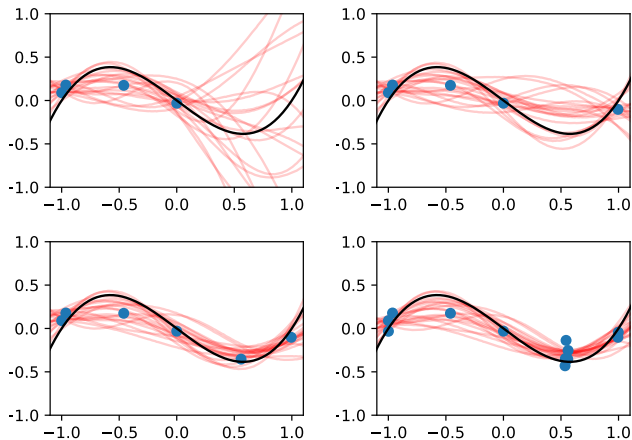
$$\begin{aligned} A_{\text{EI}}(\boldsymbol{\theta}) &:= \mathbb{E}(\min[\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta}) - \mu_{\min}, 0]) \\ &= (\mu_{\min} - \mu(\boldsymbol{\theta}))\Phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{v(\boldsymbol{\theta})}}\right) \\ &\quad + \sqrt{v(\boldsymbol{\theta})}\phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{v(\boldsymbol{\theta})}}\right) \end{aligned}$$

- $\mu_{\min} := \min_{\boldsymbol{\theta}} \mu(\boldsymbol{\theta})$
- $\Phi, \phi$  CDF and PDF of standard normal

# Bayesian Acquisition

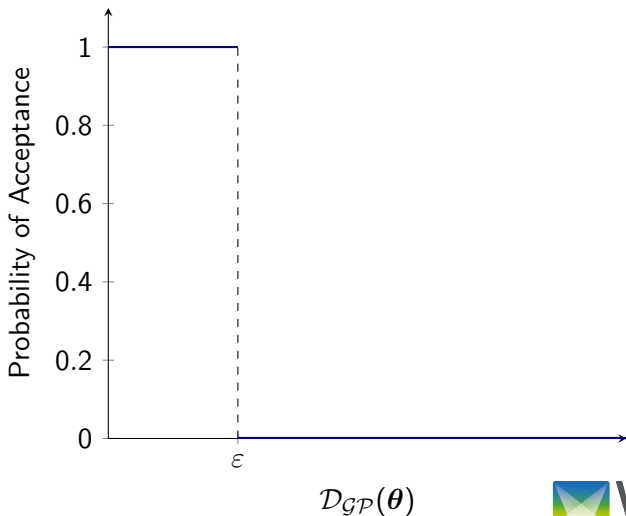
- ▶ Theoretical guarantees highly sensitive to choice of kernel

# Lower Confidence Bound



# Synthetic Likelihood

- $L(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) \approx P(\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta}) < \varepsilon)$  (up to a proportion)



# Vivax Malaria

- ▶ Has dormant liver stage on top of blood stage infection that can cause relapse.

# Vivax Model - Champagne et. al

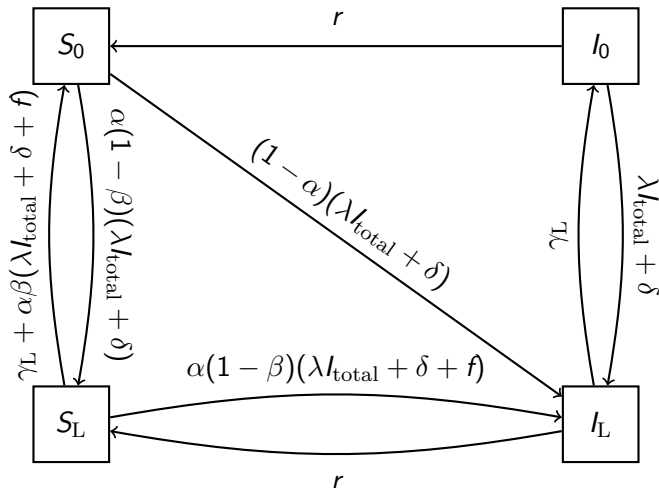


Figure: *P. vivax* model described by Champagne et al. 2022

# Ordinary Differential Equations - Champagne et. al

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

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

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

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



# Champagne Model Parameters

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

# Model Calibration Data

- ▶ Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.



Expected information acquisition function



# Model Calibration Data

- ▶ Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.
- ▶  $S(\mathbf{Y}_{\text{obs}}) := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$ 
  - ▶  $w_{\text{obs}}$  : weekly incidence around (stochastic) equilibrium
  - ▶  $p_{\text{obs}}$  : prevalence around (stochastic) equilibrium
  - ▶  $m_{\text{obs}}$  : incidence in the first month of the epidemic



Expected information acquisition function



# Model Calibration Data

- ▶ Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.
- ▶  $S(\mathbf{Y}_{\text{obs}}) := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$ 
  - ▶  $w_{\text{obs}}$  : weekly incidence around (stochastic) equilibrium
  - ▶  $p_{\text{obs}}$  : prevalence around (stochastic) equilibrium
  - ▶  $m_{\text{obs}}$  : incidence in the first month of the epidemic
- ▶

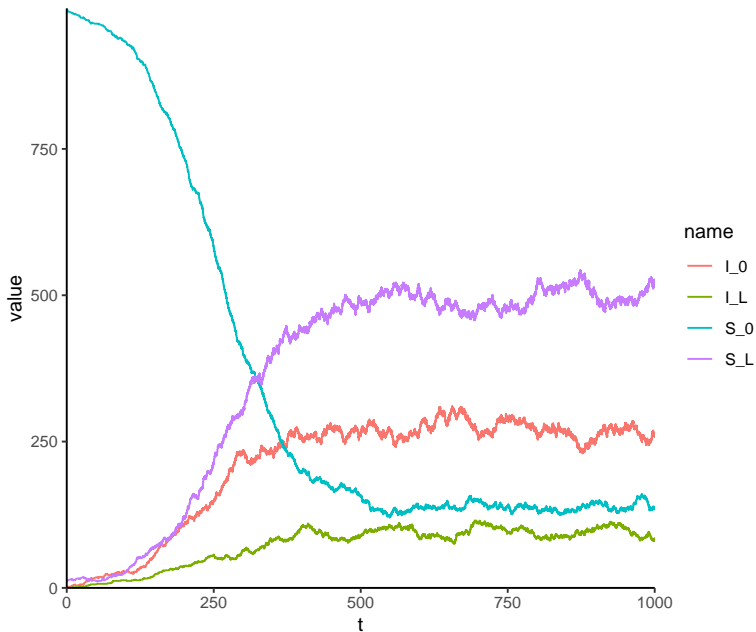
$$\mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r) = \ln \sqrt{\left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2 + \left(\frac{m - m_{\text{obs}}}{m_{\text{obs}}}\right)^2 + \left(\frac{w - w_{\text{obs}}}{w_{\text{obs}}}\right)^2}$$

- ▶ (Log of the  $L_2$  norm of the relative differences)

Expected information acquisition function



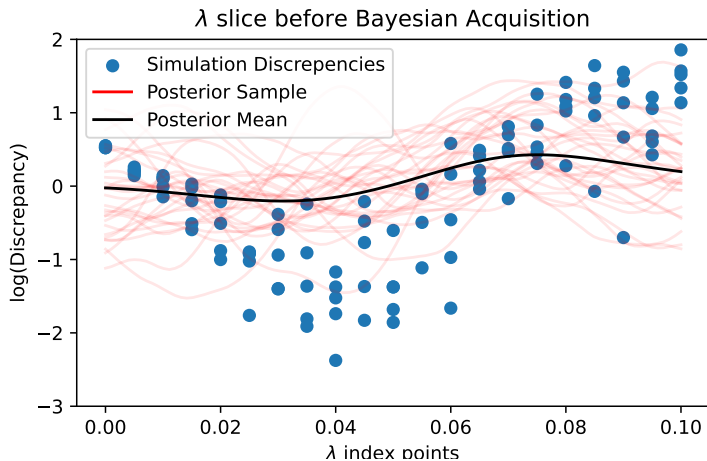
# Example Simulation



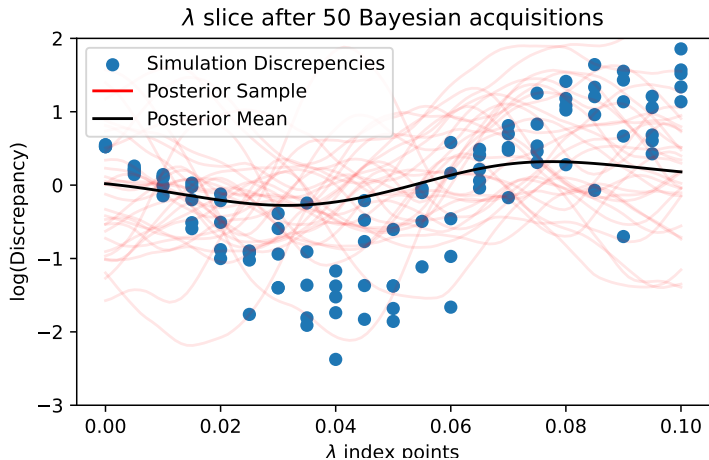
# How did it go?



# How did it go?

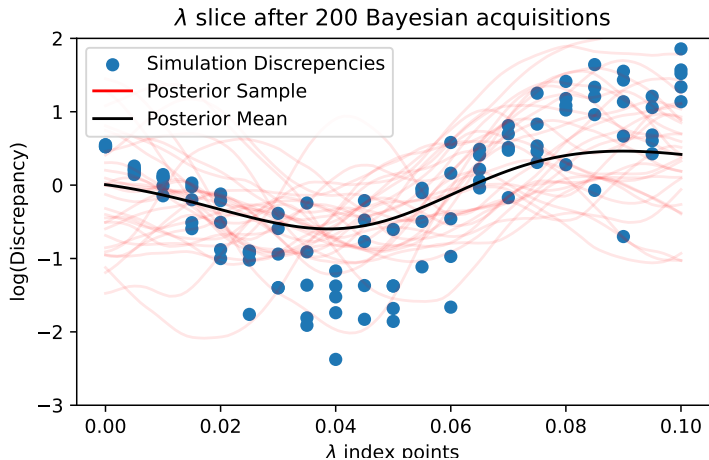


# How did it go?

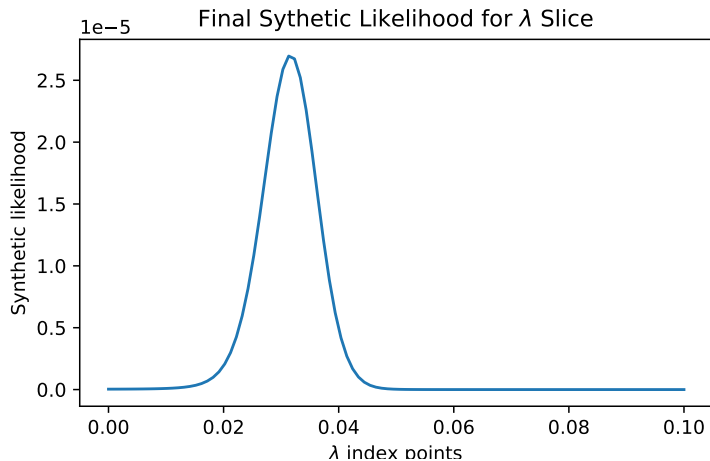




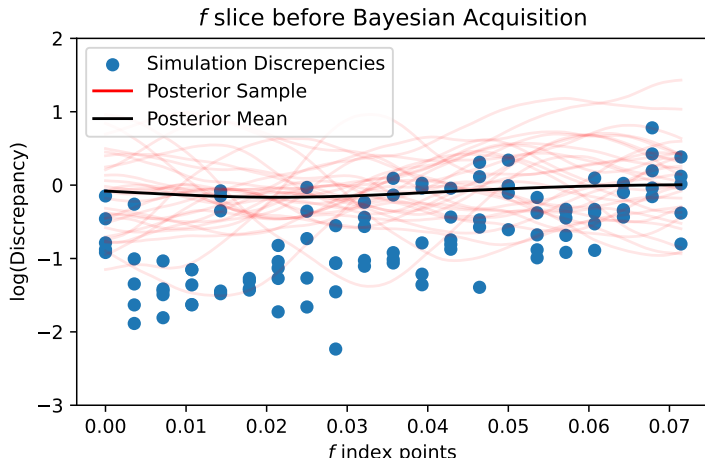
# How did it go?



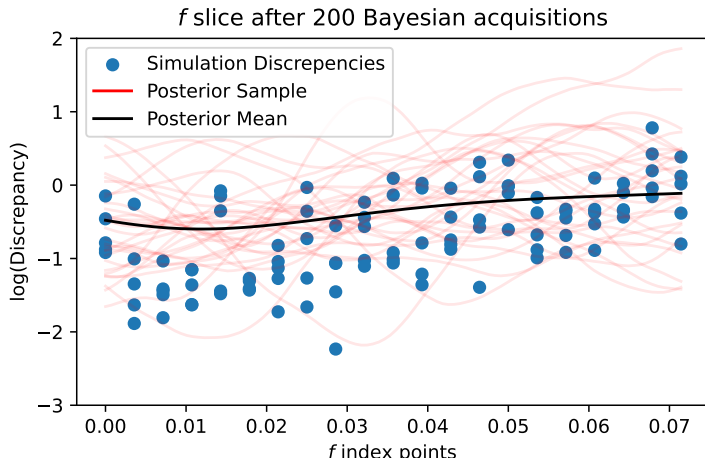
# How did it go?



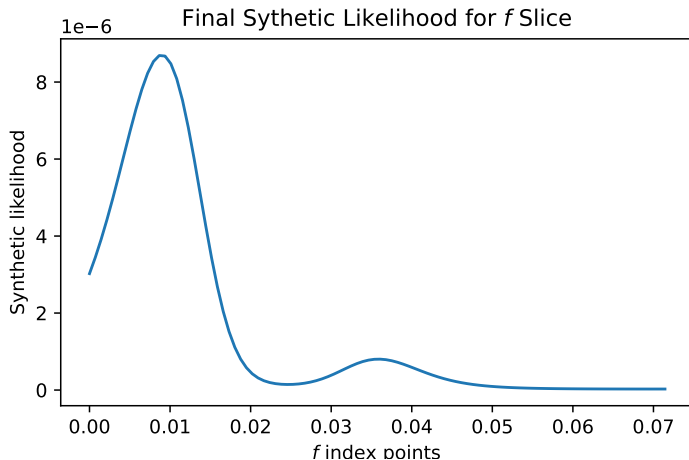
# How did it go?



# 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

