

# Efficient Likelihood Approximation via Gaussian Processes

With an Application to a *P. Vivax* Malaria Model

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

June 2024



# Introduction and Motivation

- ▶ 600,000 deaths/year, 75% children under 5
- ▶ Two main species *P. vivax* and *P. falciparum*
- ▶ *P. falciparum* main cause of death, but *P. vivax* historically underestimated.
- ▶ Proportion of *P. vivax* cases increased over last 50 years.

# *P. vivax* has Dormant Stage

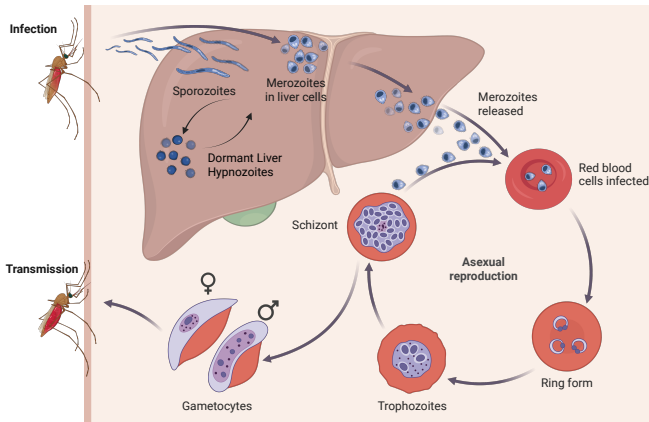


Figure: *P. vivax* lifecycle. Created with BioRender.com

# Vivax Model - Champagne et. al

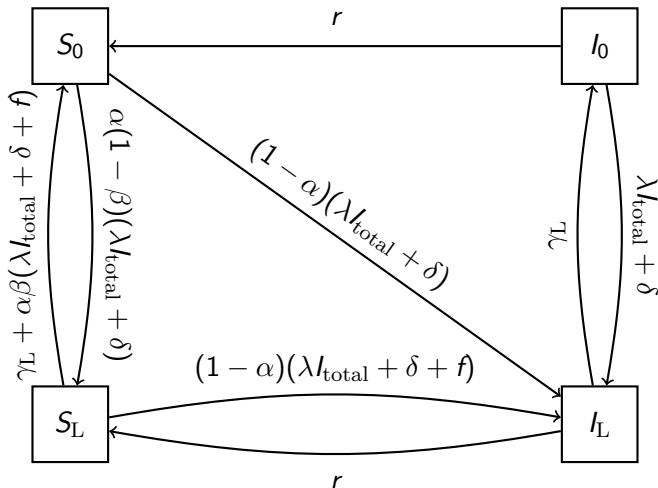


Figure: Champagne et al. 2022 *P. vivax* model

# Champagne Model Parameters

- ▶  $\alpha$  : proportion of those infected who clear blood stage infections through treatment
- ▶  $\beta$  : proportion of those cleared of blood stage infection who are also cleared of liver stage parasites (radical cure)
- ▶  $\lambda$  : rate of infection
- ▶  $\gamma_L$  : rate of liver stage disease clearance
- ▶  $f$  : rate of relapse
- ▶  $r$  : rate of blood stage clearance
- ▶  $\delta = 0$  importation rate (fixed)

# The Problem

- ▶ How to calibrate model parameters?

# The Problem

- ▶ How to calibrate model parameters?
- ▶ Simulations take long time (and models get a lot more complicated)

# Notation

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



# Notation

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

# In an ideal world...

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

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

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

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  $\theta_i$  from prior
2. Run model
3. Accept  $\theta_i$  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  $\theta$  is near  $\theta'$ .
  - ▶ Gaussian Processes

# Gaussian Process Setup

Formally we can assume that

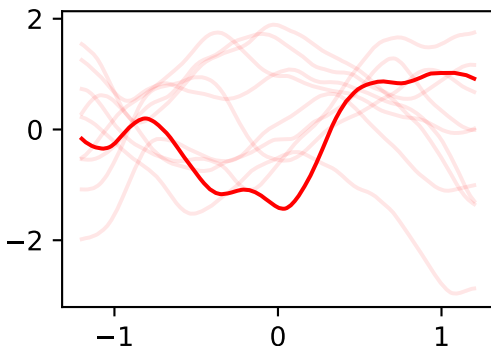
$$\text{Cov}(\mathcal{D}(\boldsymbol{\theta}_i), \mathcal{D}(\boldsymbol{\theta}_j)) = k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$$

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.*



# 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

- ▶ Choice of kernel determines smoothness
- ▶ 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

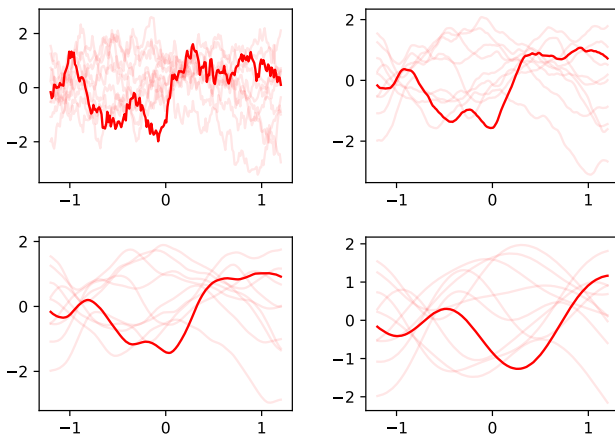


Figure: Matérn 1/2, 3/2, 5/2, and squared exponential kernels.



# 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
- ▶  $\sigma_k^2 = \text{Var}(f(\mathbf{x}))$

# Discrepancy Function Context

- ▶ Long term play: replace  $\mathcal{D}(\theta)$  with a Gaussian process surrogate model approximation.

# Discrepancy Function Context

- ▶ Long term play: replace  $\mathcal{D}(\theta)$  with a Gaussian process surrogate model approximation.
- ▶ What if we have observations already?

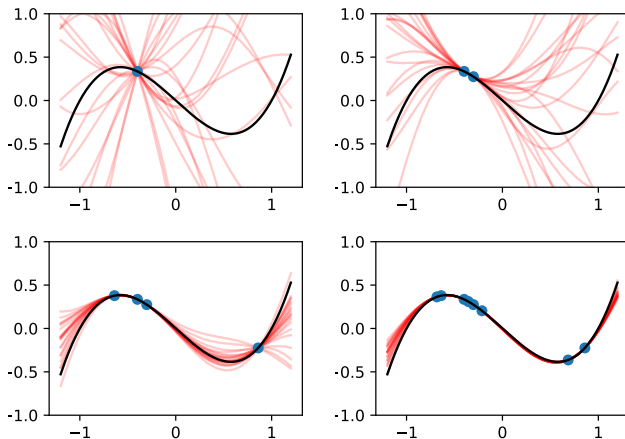
# Gaussian Process Regression

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

implies

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

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

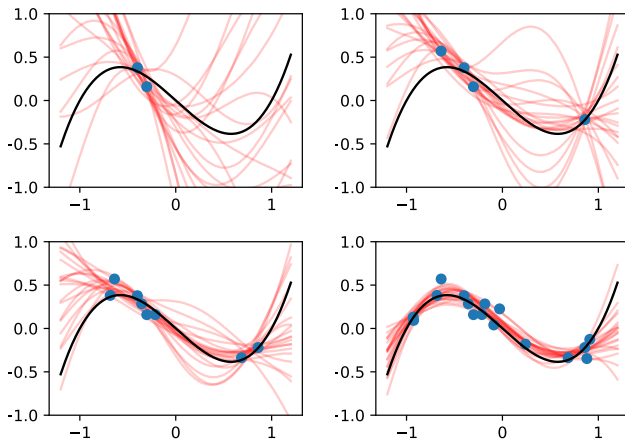


# Normal observation noise

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 Idea

- ▶ If  $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$  can be well approximated by a Gaussian process and ...



# Key Idea

- ▶ If  $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$  can be well approximated by a Gaussian process and ...
- ▶  $\mathcal{D}(\boldsymbol{\theta})$  approximately distributed  $N(\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})], \sigma_o^2)$  then...

# Key Idea

- ▶ If  $\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})]$  can be well approximated by a Gaussian process and ...
- ▶  $\mathcal{D}(\boldsymbol{\theta})$  approximately distributed  $N(\mathbb{E}[\mathcal{D}(\boldsymbol{\theta})], \sigma_o^2)$  then...
- ▶ Approximate  $\mathcal{D}(\boldsymbol{\theta})$  with  $\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta})$ , a Gaussian process with observation noise.

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

# Approximate ABC...??

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

# Synthetic Likelihood

Pr drawing and accepting  $\theta$  using 'approximate' ABC is

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

and hence

$$\hat{L}(\theta) := \Pr(\mathcal{D}_{\mathcal{GP}}(\theta) < \varepsilon) \approx c\mathcal{L}(\theta)$$

for some  $c$ .

# 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 becomes:
  - ▶  $\mathcal{D}(\boldsymbol{\theta})$  approximately distributed  $LN(\cdot, \sigma_o^2)$ .

# Log Gaussian Process

- ▶ Alternatively we can model  $\ln \mathcal{D}(\boldsymbol{\theta})$  as a Gaussian process  $d_{\mathcal{GP}}(\boldsymbol{\theta})$ .
- ▶ Key assumptions becomes:
  - ▶  $\mathcal{D}(\boldsymbol{\theta})$  approximately distributed  $LN(\cdot, \sigma_o^2)$ .



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



# 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:
  - ▶  $\mathbb{E}[\mathcal{D}(\theta)]$  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:
  - ▶  $\mathbb{E}[\mathcal{D}(\theta)]$  small,
  - ▶  $\mathcal{D}(\theta)$  highly unknown.

# Bayesian Acquisition Functions

- ▶  $\mu(\boldsymbol{\theta}) := \mathbb{E}(D_{\mathcal{GP}}(\boldsymbol{\theta}))$  and  $v(\boldsymbol{\theta}) := \text{Var}(D_{\mathcal{GP}}(\boldsymbol{\theta}))$
- ▶ Bayesian acquisition functions  $A(\boldsymbol{\theta})$ , quantify how 'desirable' it is to sample  $\mathcal{D}(\boldsymbol{\theta})$  at  $\boldsymbol{\theta}$ .

# Bayesian Acquisition Functions

- ▶  $\mu(\boldsymbol{\theta}) := \mathbb{E}(D_{\mathcal{GP}}(\boldsymbol{\theta}))$  and  $v(\boldsymbol{\theta}) := \text{Var}(D_{\mathcal{GP}}(\boldsymbol{\theta}))$
- ▶ Bayesian acquisition functions  $A(\boldsymbol{\theta})$ , quantify how 'desirable' it is to sample  $\mathcal{D}(\boldsymbol{\theta})$  at  $\boldsymbol{\theta}$ .
- ▶ Gutmann and Cor 2016 minimise lower confidence bound

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

the lower value of a confidence interval (determined by  $\eta_t$ ).

# Vivax Model - Champagne et. al

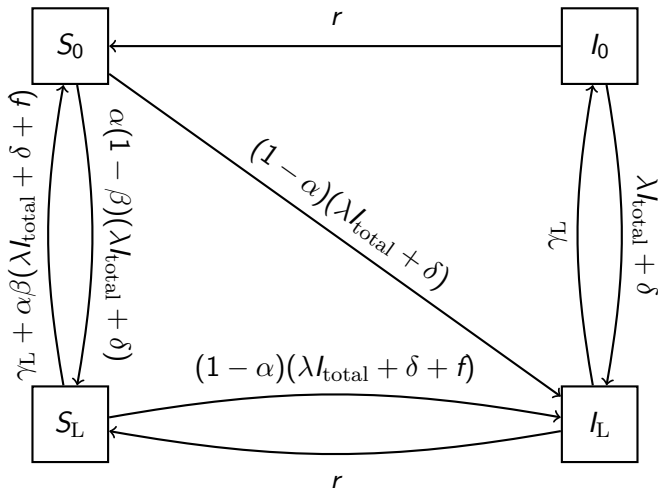
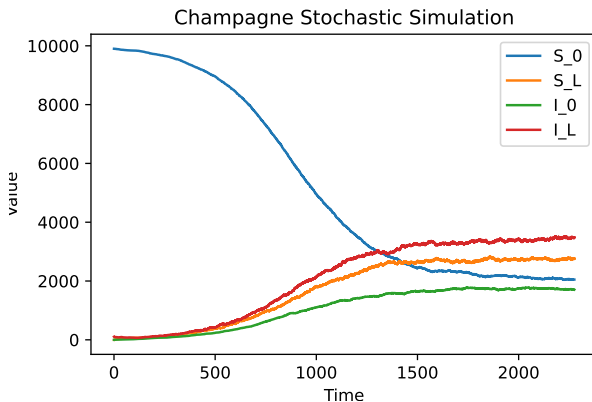


Figure: Champagne et al. 2022 *P. vivax* model

# 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

# Model Simulation



**Figure:** Exact stochastic simulation using parameters reported in Champagne et al. 2022. Population 10,000, initial infections 100.



# 'Observed' Data

- ▶  $\mathbf{Y}_{\text{obs}} := \{\iota_{\text{obs}}, \pi_{\text{obs}}, i_{\text{obs}}, p_{\text{obs}}\}$ 
  - ▶  $\iota_{\text{obs}}$  : weekly incidence around steady state equilibrium
  - ▶  $\pi_{\text{obs}}$  : prevalence around steady state equilibrium
  - ▶  $i_{\text{obs}}$  : incidence in the first month of the epidemic
  - ▶  $p_{\text{obs}}$  : prevalence after one month of simulation

# 'Observed' Data

- ▶  $\mathbf{Y}_{\text{obs}} := \{\iota_{\text{obs}}, \pi_{\text{obs}}, i_{\text{obs}}, p_{\text{obs}}\}$ 
  - ▶  $\iota_{\text{obs}}$  : weekly incidence around steady state equilibrium
  - ▶  $\pi_{\text{obs}}$  : prevalence around steady state equilibrium
  - ▶  $i_{\text{obs}}$  : incidence in the first month of the epidemic
  - ▶  $p_{\text{obs}}$  : prevalence after one month of simulation
- ▶  $\mathcal{D}(\alpha, \beta, \gamma_L, \lambda, f, r)$  is the  $L_2$  norm of the relative differences

$$\sqrt{\left(\frac{\iota - \iota_{\text{obs}}}{\iota_{\text{obs}}}\right)^2 + \left(\frac{\pi - \pi_{\text{obs}}}{\pi_{\text{obs}}}\right)^2 + \left(\frac{i - i_{\text{obs}}}{i_{\text{obs}}}\right)^2 + \left(\frac{p - p_{\text{obs}}}{p_{\text{obs}}}\right)^2}$$

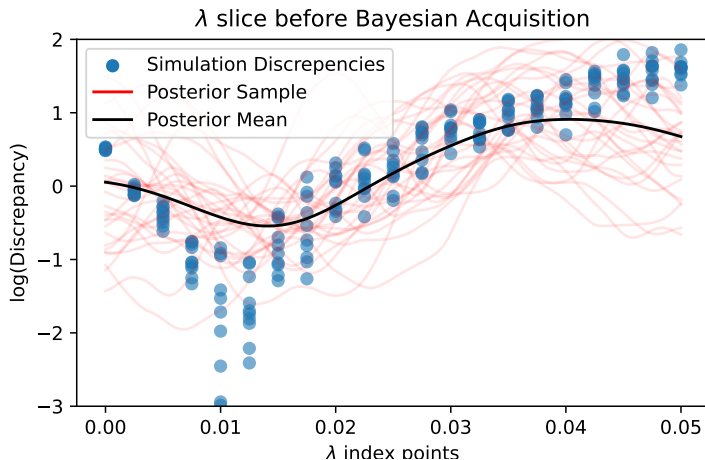
# GP choices

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

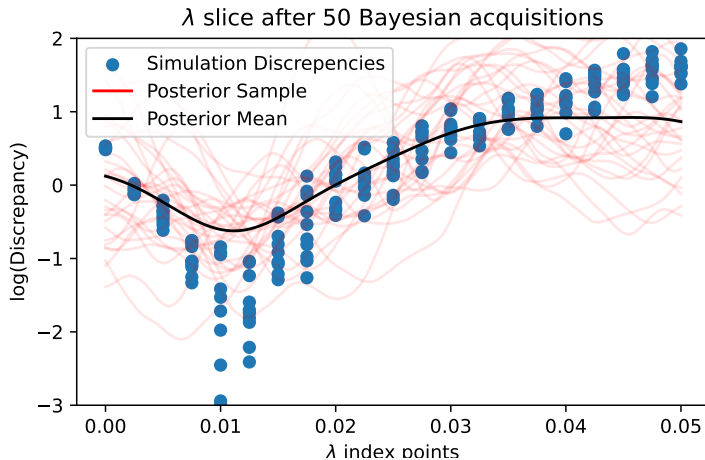
# How did it go?



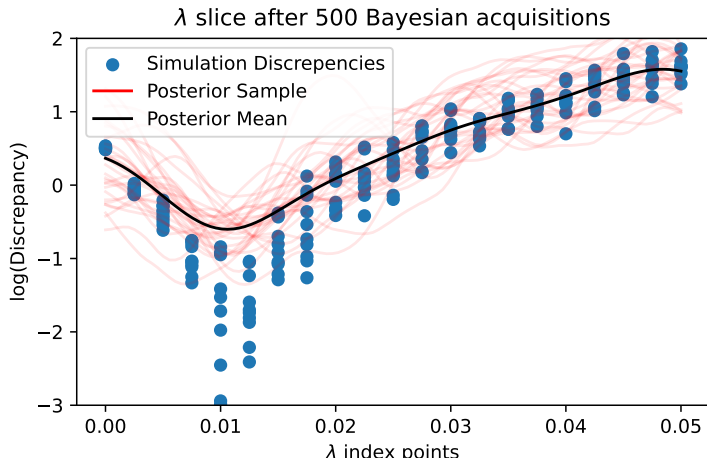
# How did it go?



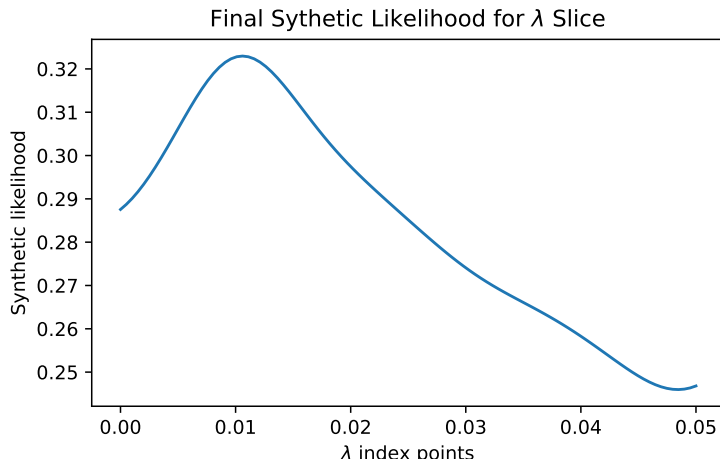
# How did it go?



# How did it go?

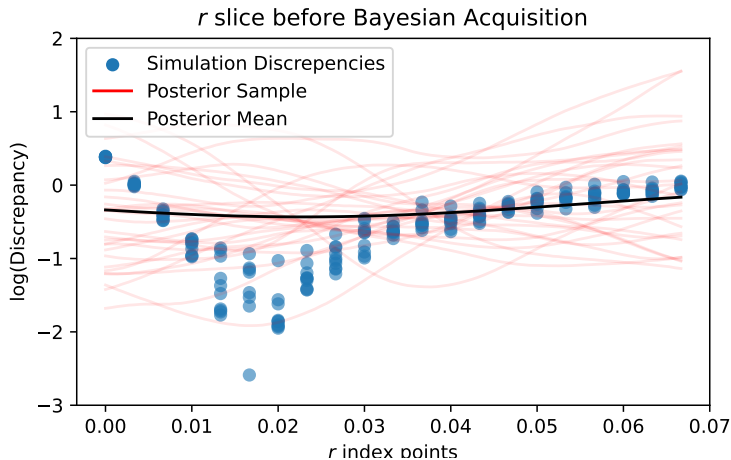


# How did it go?

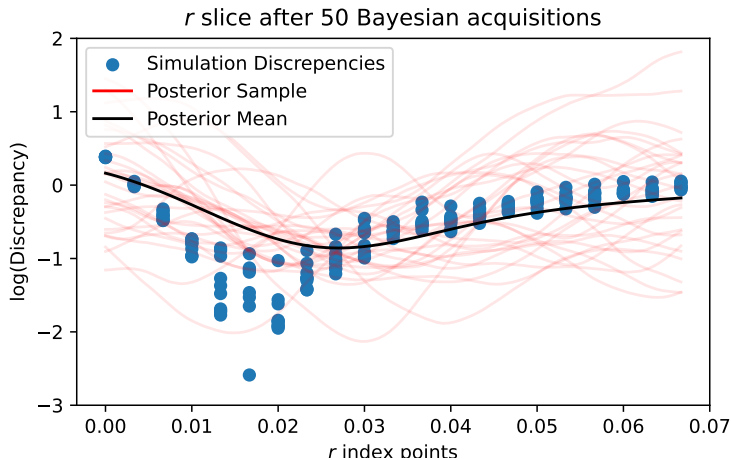




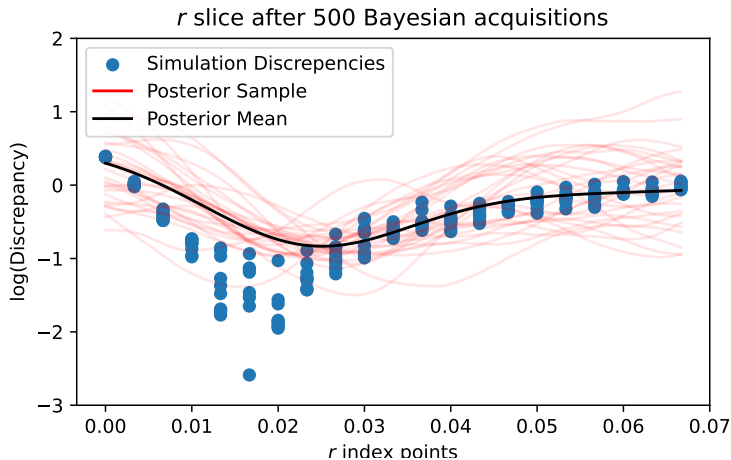
# How did it go?



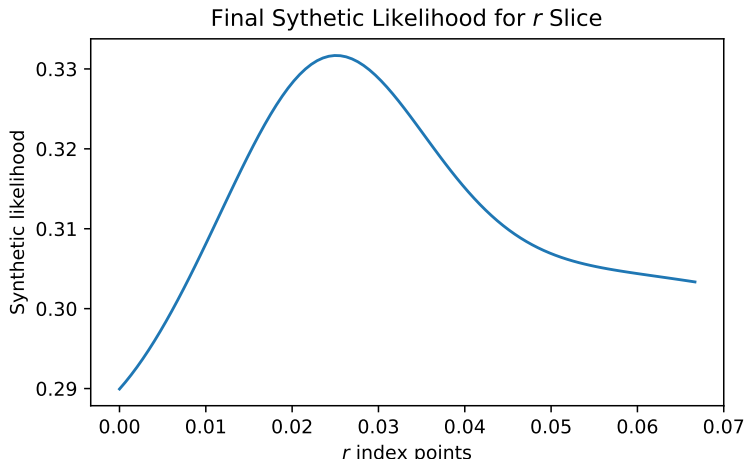
# How did it go?



# How did it go?



# How did it go?



# Discussion

- ▶ Bifurcation points effect:
  1. observation variance,
  2. distribution of observations,
  3. behaviour of the discrepancy mean.
- ▶ Possible extensions:

# Discussion

- ▶ Bifurcation points effect:
  1. observation variance,
  2. distribution of observations,
  3. behaviour of the discrepancy mean.
- ▶ Possible extensions:
  1. model  $s^2(\theta)$ ,

# Discussion

- ▶ Bifurcation points effect:
  1. observation variance,
  2. distribution of observations,
  3. behaviour of the discrepancy mean.
- ▶ Possible extensions:
  1. model  $s^2(\theta)$ ,
  2. choose a different distribution and moment match,

# Discussion

- ▶ Bifurcation points effect:
  1. observation variance,
  2. distribution of observations,
  3. behaviour of the discrepancy mean.
- ▶ Possible extensions:
  1. model  $s^2(\theta)$ ,
  2. choose a different distribution and moment match,
  3. use a Student's  $t$ -process.



# Conclusion

- ▶ Calibrating model parameters is important for scenario testing etc
- ▶ Successfully calibrated model parameters
- ▶ Could be used with more complicated models, even your model...

# Thanks to

- ▶ Eamon Conway
- ▶ Jennifer Flegg
- ▶ Ivo Mueller
- ▶ Mueller lab and unimelb MMB group



# 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).