

# Bayesian Optimisation for Likelihood Free Inference

Make model parameterisation go brrr

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

April 2024



# Notation

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



# Notation

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



# Notation

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



# 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
- ▶  $\delta$  : importation rate (which we assume is 0)



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

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



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

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



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

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



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

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



# Reality Sets In

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



# A Standard Bayesian Solution

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



# What is 'matches'

- ▶ Match if  $\mathbf{X}_\theta = \mathbf{X}_{\text{obs}}$ ? (no)



# What is 'matches'

- ▶ Match if  $\mathbf{X}_\theta = \mathbf{X}_{\text{obs}}$ ? (no)
  - ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
    - ▶ Can be a norm such as
- $$\|S(\mathbf{X}_\theta) - S(\mathbf{X}_{\text{obs}})\|_p := (\sum_{i=1}^d |S(\mathbf{X}_\theta)_i - S(\mathbf{X}_{\text{obs}})_i|^p)^{1/p}$$



# What is 'matches'

- ▶ Match if  $\mathbf{X}_\theta = \mathbf{X}_{\text{obs}}$ ? (no)
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ Can be a norm such as
$$\|S(\mathbf{X}_\theta) - S(\mathbf{X}_{\text{obs}})\|_p := (\sum_{i=1}^d |S(\mathbf{X}_\theta)_i - S(\mathbf{X}_{\text{obs}})_i|^p)^{1/p}$$
  - ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).



# What is 'matches'

- ▶ Match if  $\mathbf{X}_\theta = \mathbf{X}_{\text{obs}}$ ? (no)
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ Can be a norm such as
$$\|S(\mathbf{X}_\theta) - S(\mathbf{X}_{\text{obs}})\|_p := (\sum_{i=1}^d |S(\mathbf{X}_\theta)_i - S(\mathbf{X}_{\text{obs}})_i|^p)^{1/p}$$
  - ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).
- ▶ We consider  $\mathcal{D}(\theta) := D(S(\mathbf{X}_\theta), S(\mathbf{X}_{\text{obs}}))$

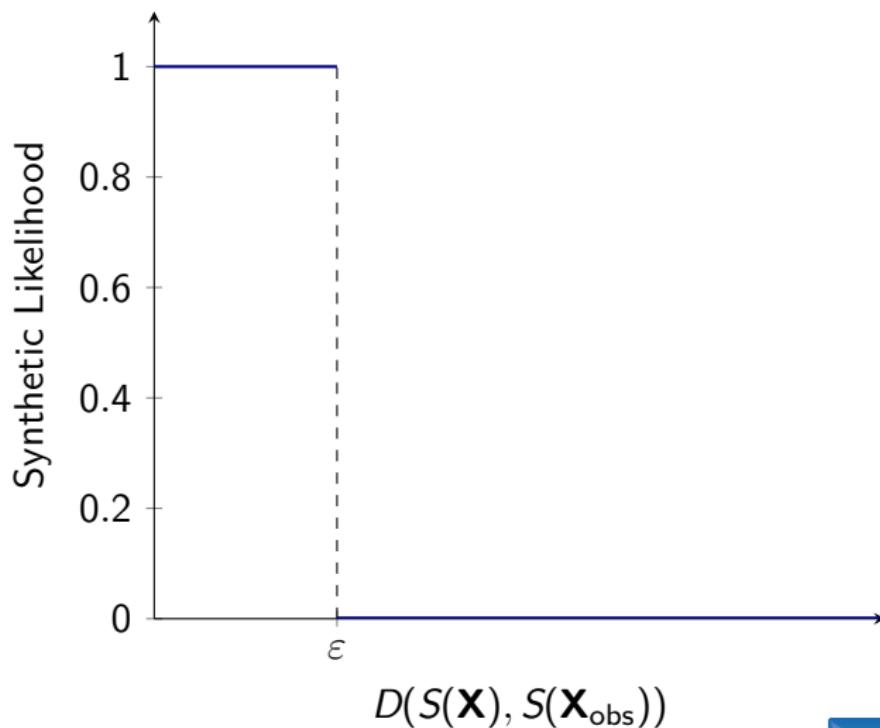


# What is 'matches'

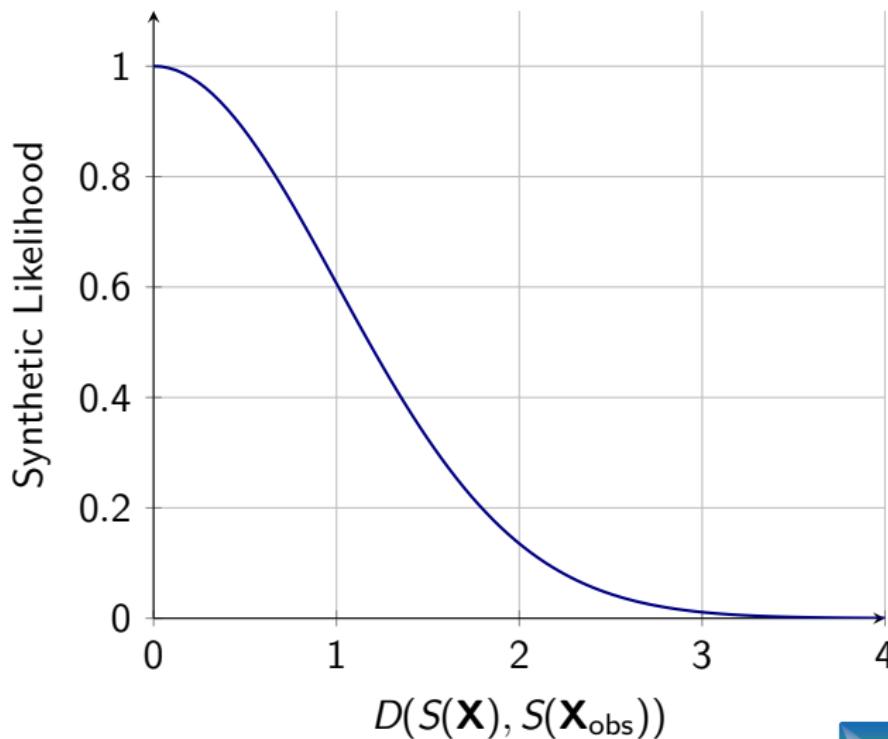
- ▶ Match if  $\mathbf{X}_\theta = \mathbf{X}_{\text{obs}}$ ? (no)
- ▶ Discrepancy function  $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ 
  - ▶ Can be a norm such as
$$\|S(\mathbf{X}_\theta) - S(\mathbf{X}_{\text{obs}})\|_p := (\sum_{i=1}^d |S(\mathbf{X}_\theta)_i - S(\mathbf{X}_{\text{obs}})_i|^p)^{1/p}$$
  - ▶ Rescale  $S(\cdot)$  appropriately (ie via a covariance matrix).
- ▶ We consider  $\mathcal{D}(\theta) := D(S(\mathbf{X}_\theta), S(\mathbf{X}_{\text{obs}}))$
- ▶  $\mathcal{D}(\theta)$  is 'how close' we were using parameters  $\theta$ .



# Uniform Acceptance Probability



# Acceptance Probability



# Overall Idea of my Research

- ▶ Can we predict  $\mathcal{D}(\theta)$  for simulated  $\theta$ ?



# Overall Idea of my Research

- ▶ Can we predict  $\mathcal{D}(\theta)$  for simulated  $\theta$ ?
- ▶ Locally, hopefully yes.



# Gaussian Processes

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



# Gaussian Processes

- ▶ Random functions
- ▶ Common examples - Brownian motion, Ornstein Uhlenbeck process
- ▶ Nap time for wet-lab people



# 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  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  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

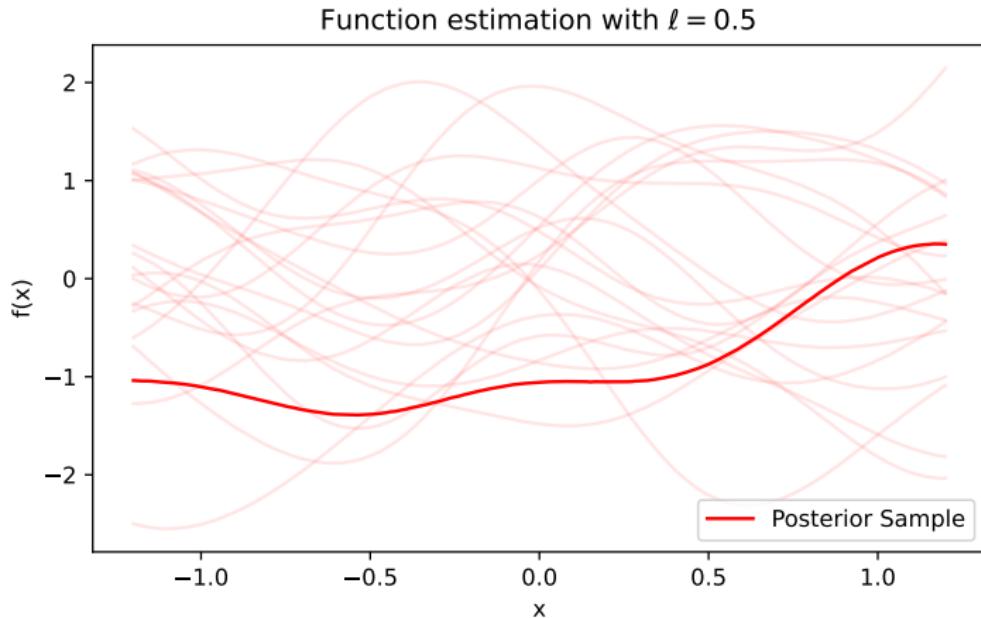
$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$



THE UNIVERSITY OF  
MELBOURNE



# Gaussian Process Example Realisations



THE UNIVERSITY OF  
MELBOURNE



# 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

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

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

- ▶  $\lfloor \nu \rfloor$  times mean square differentiable.
- ▶ As  $\nu \rightarrow \infty$  you get squared exponential covariance function, which results in realisations that are infinitely mean square differentiable:

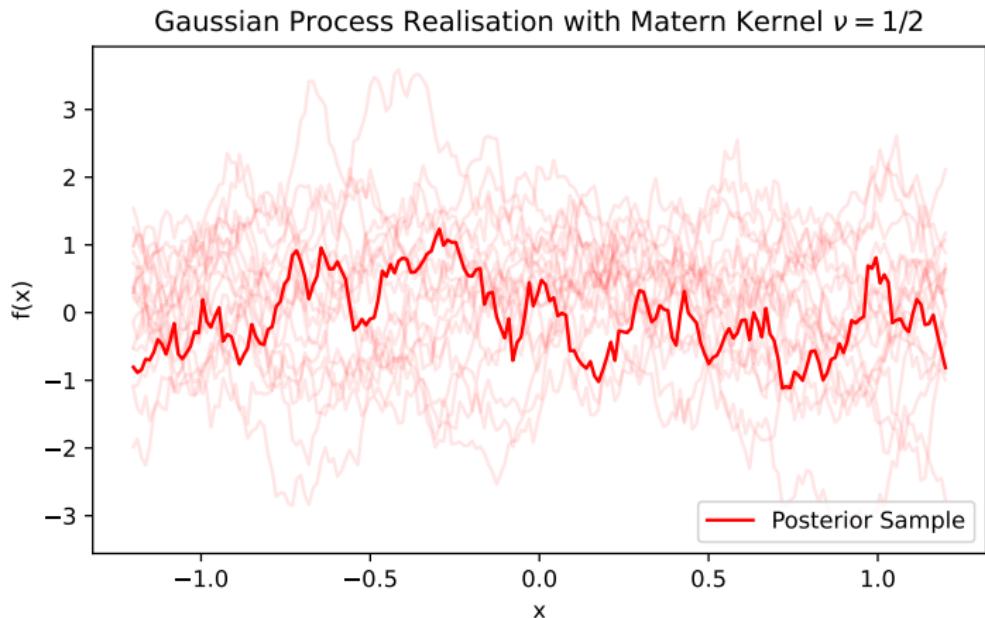
$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{\ell}\right)$$



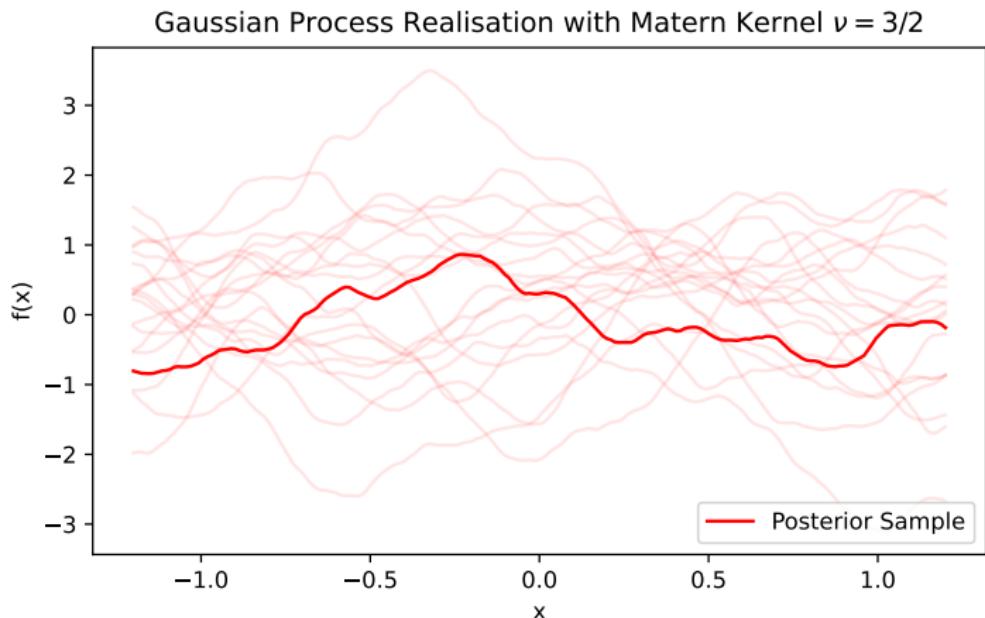
Time to wake up



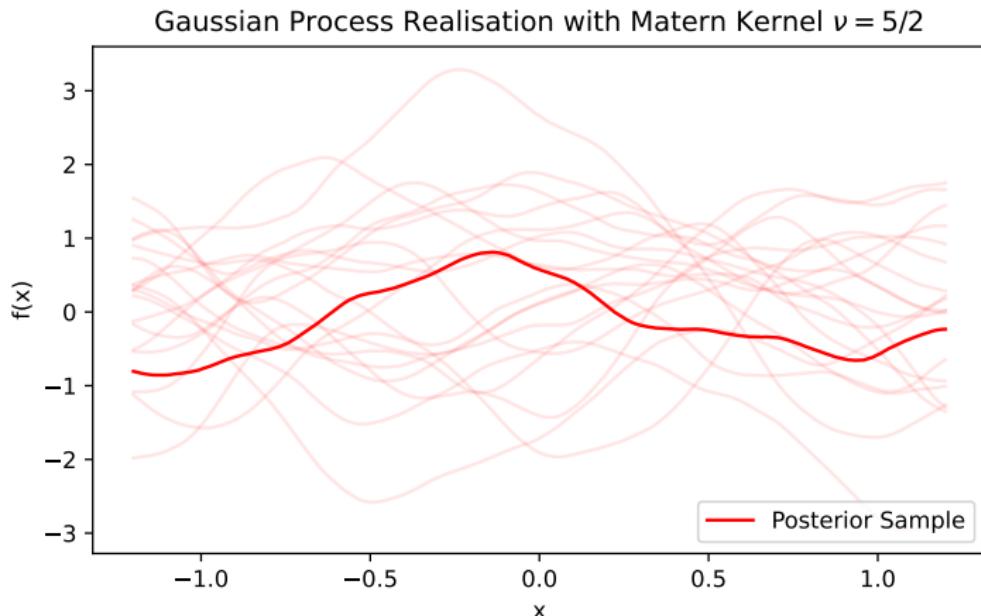
# Kernel Choices - Kernel Type



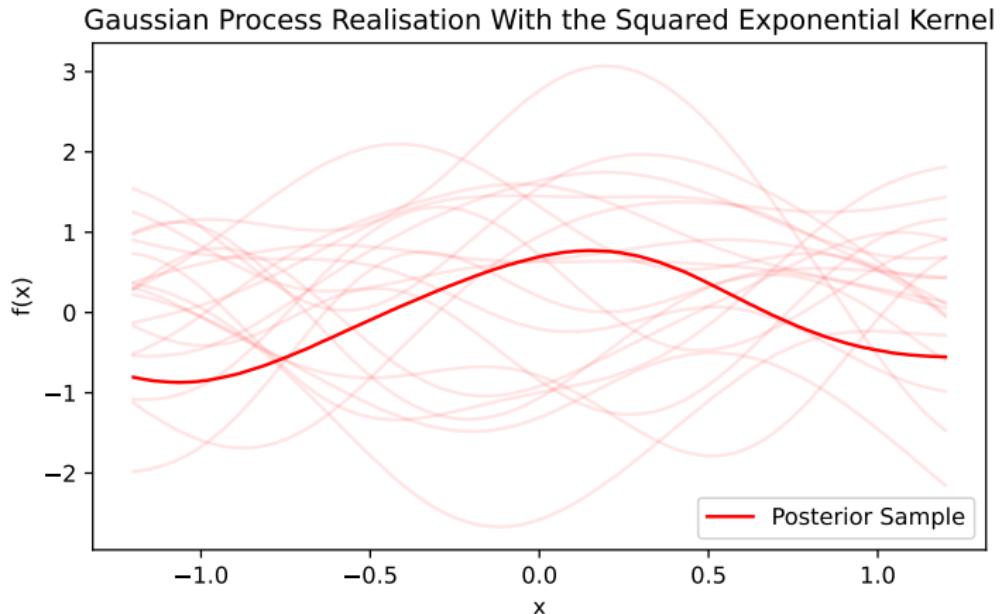
# Kernel Choices - Kernel Type



# Kernel Choices - Kernel Type



# Kernel Choices - Kernel Type



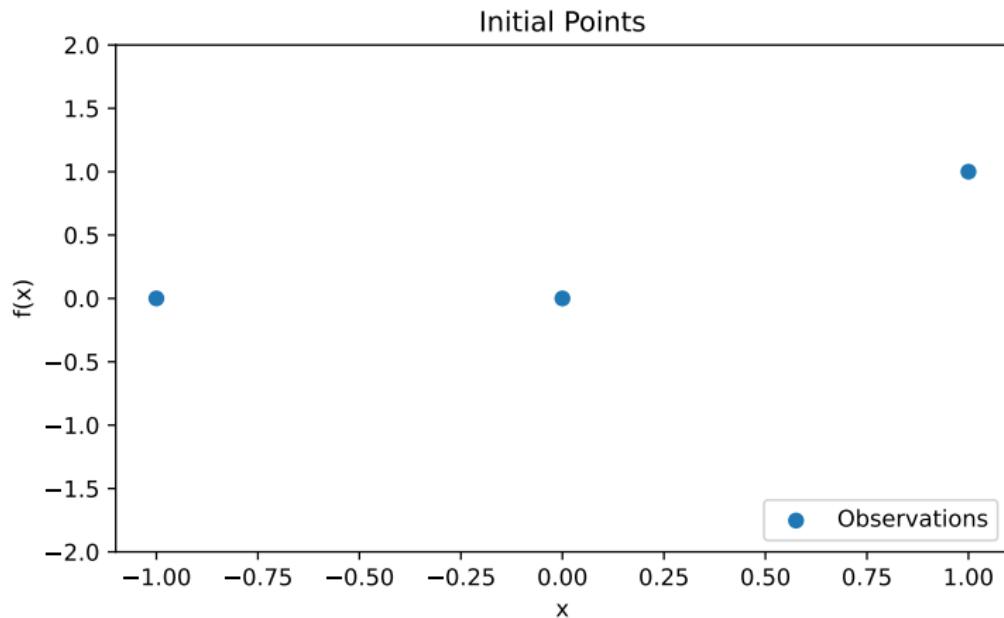
# Fitting our GP to data

GPs are 'priors'



# Fitting our GP to data

GPs are 'priors'

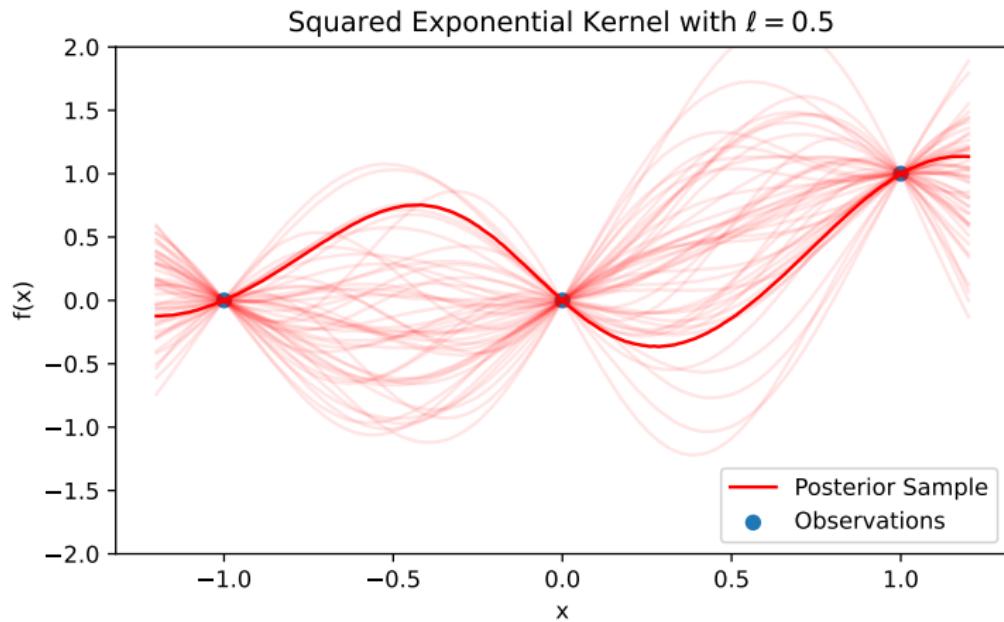


THE UNIVERSITY OF  
MELBOURNE



# Fitting our GP to data

GPs are 'priors'



# Fitting our GP to data

GPs are 'priors'



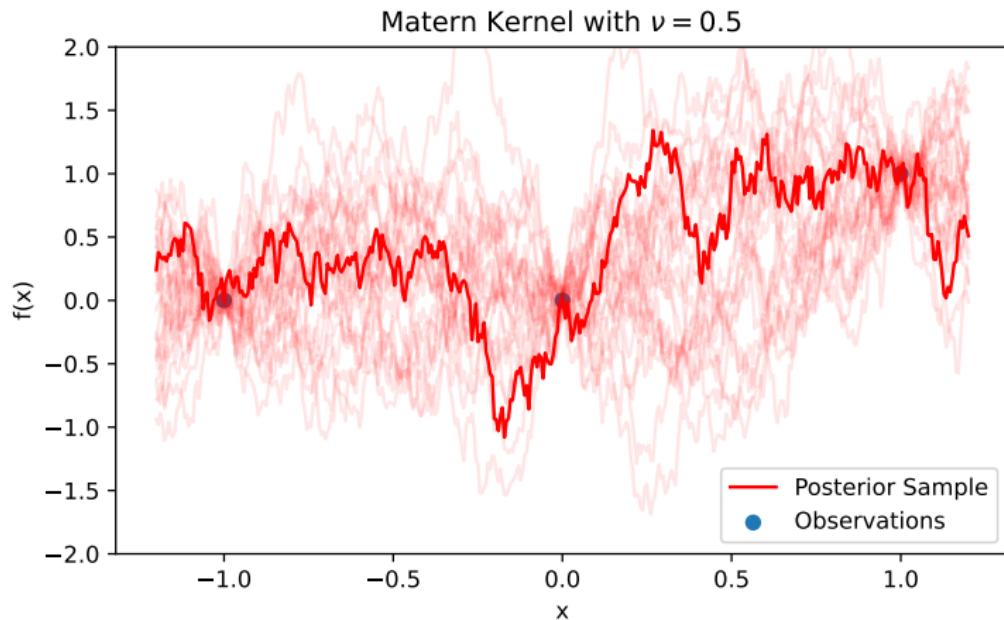
# Fitting our GP to data

GPs are 'priors'



# Fitting our GP to data

GPs are 'priors'



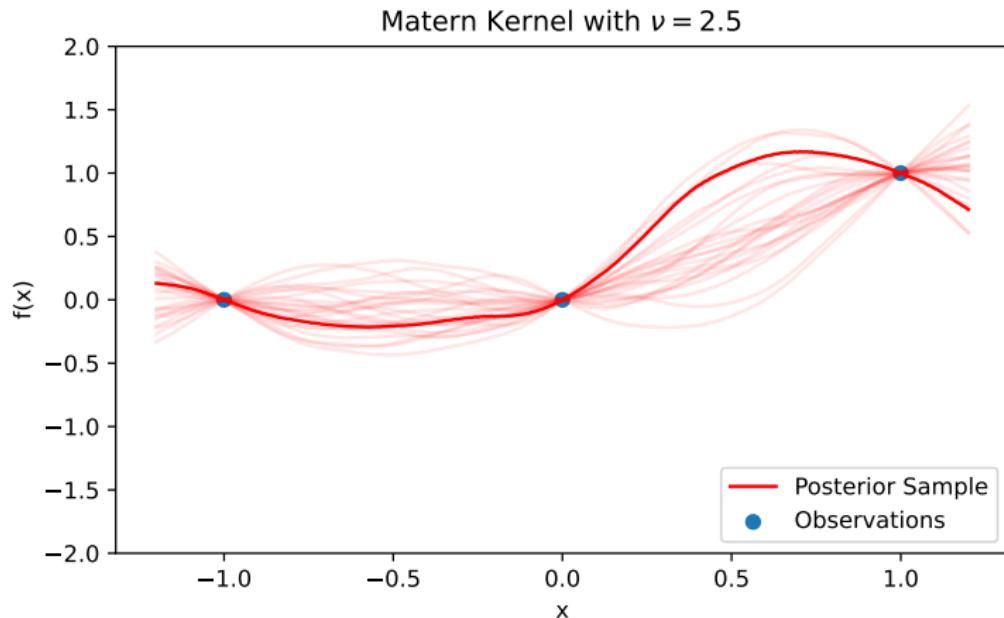
# Fitting our GP to data

GPs are 'priors'



# Fitting our GP to data

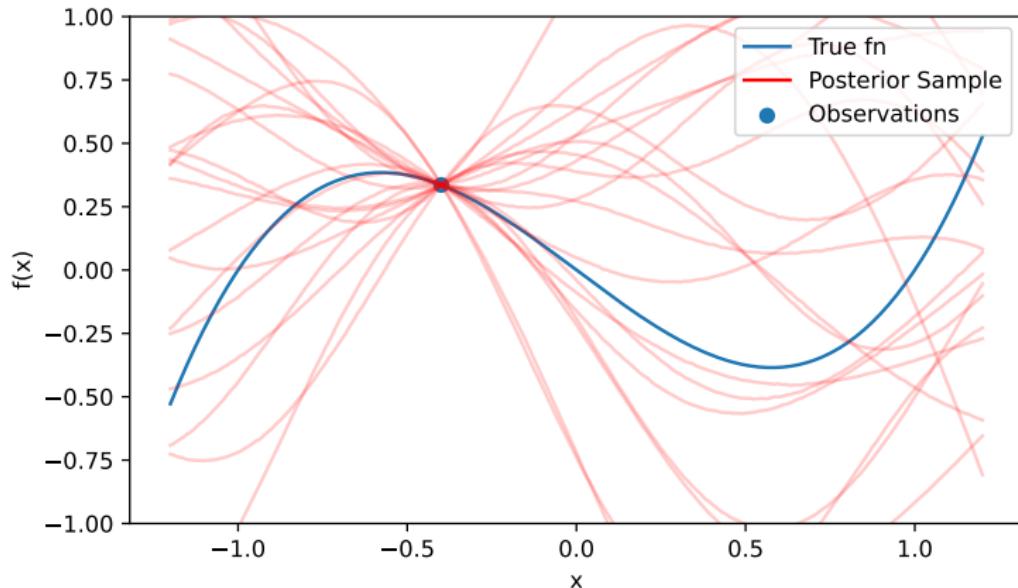
GPs are 'priors'



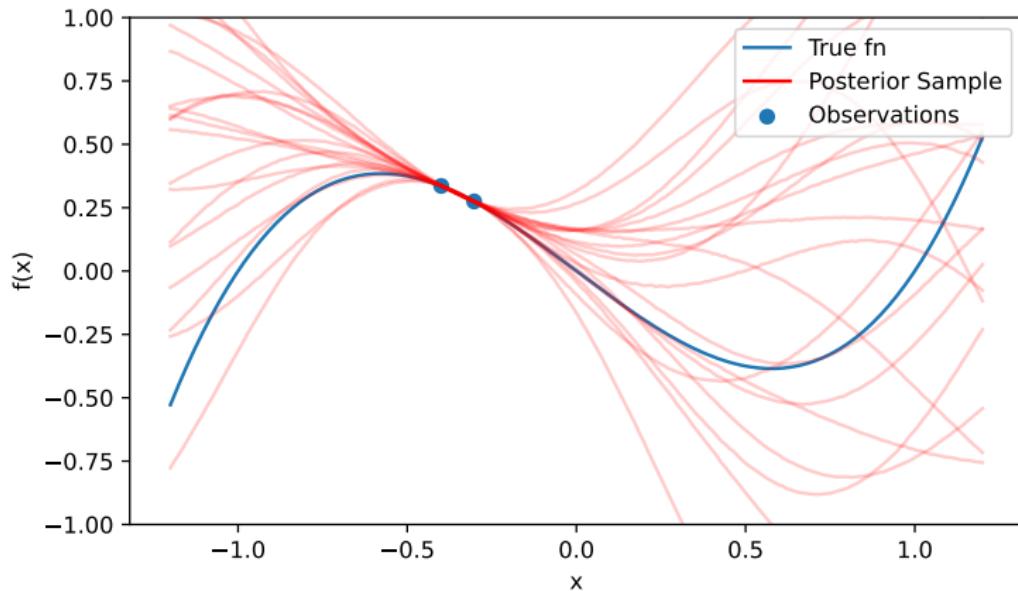
# GP regression on $x(x - 1)(x + 1)$



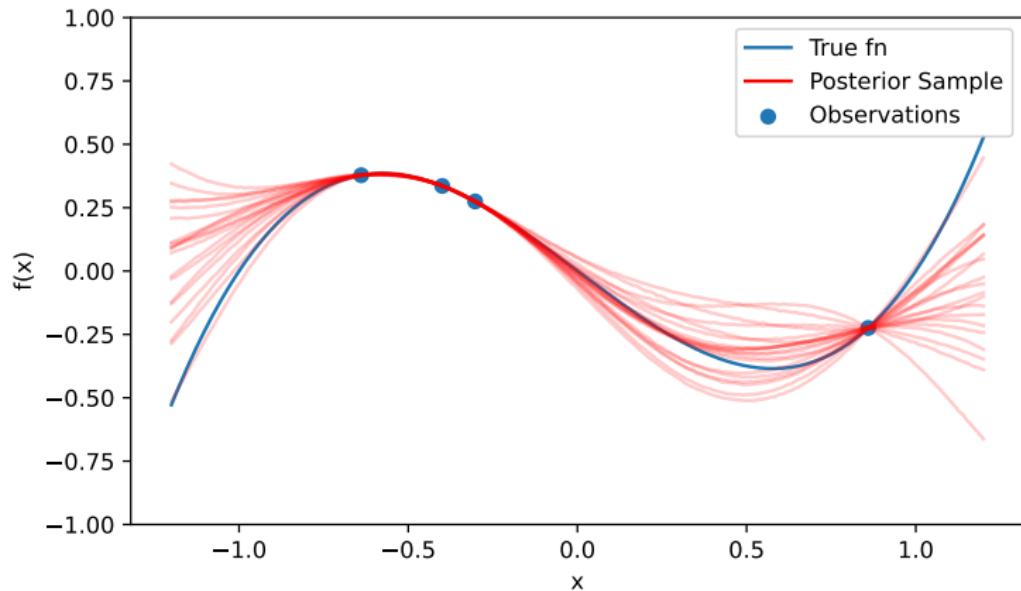
# GP regression on $x(x - 1)(x + 1)$



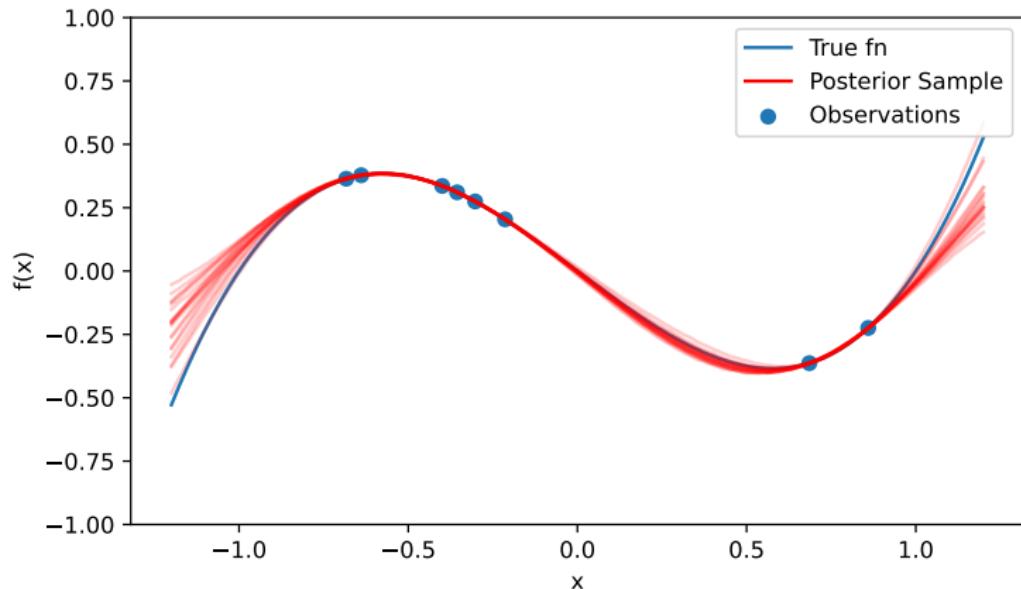
# GP regression on $x(x - 1)(x + 1)$



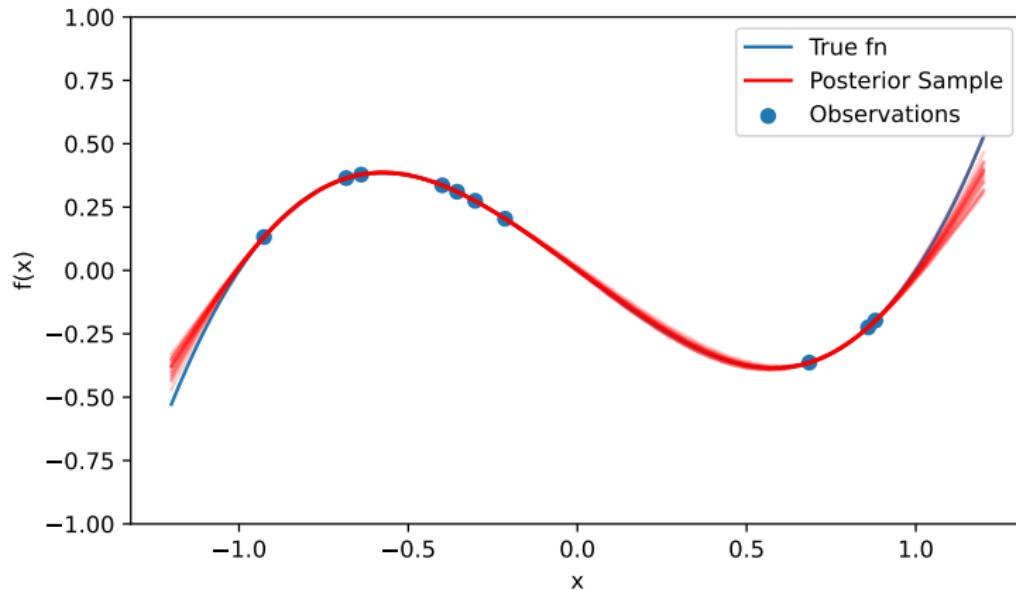
# GP regression on $x(x - 1)(x + 1)$



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



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



# What if we have noise?

Add observation variance  $\sigma^2$ ,

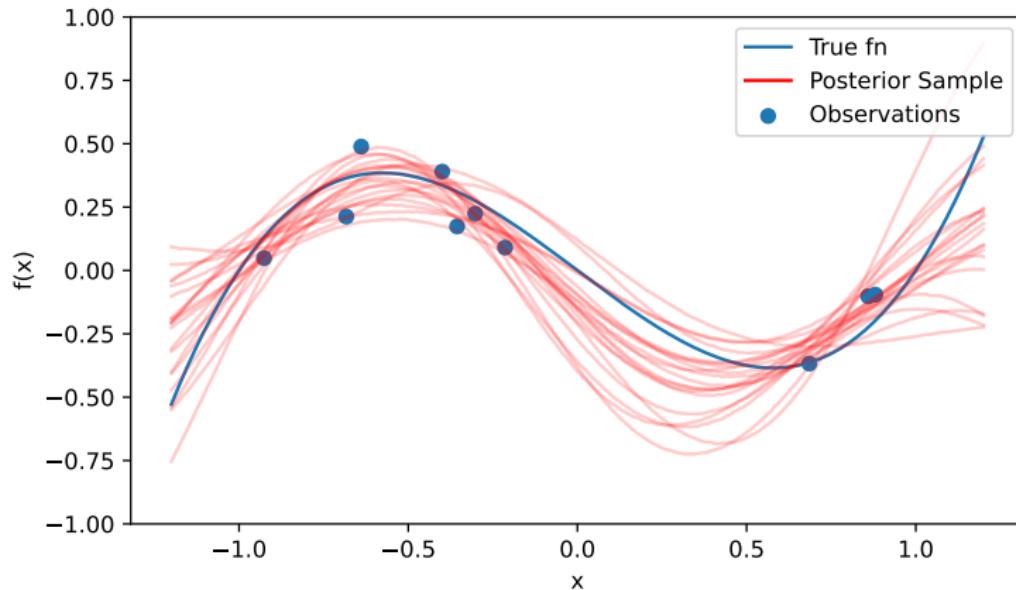
$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} + \sigma^2 \mathbf{I}_n \right)$$



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



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



# Overall Idea again

- ▶ Can we predict  $\mathcal{D}(\theta)$  for simulated  $\theta$



# Overall Idea again

- ▶ Can we predict  $\mathcal{D}(\theta)$  for simulated  $\theta$
- ▶  $\mathcal{D}(\theta) \approx \mathcal{D}(\theta')$  for  $\theta, \theta'$  close.



# Overall Idea again

- ▶ Can we predict  $\mathcal{D}(\theta)$  for simulated  $\theta$
- ▶ Use Gaussian process to predict discrepancy function



# About Vivax Malaria

- ▶ Has dormant liver stage on top of blood stage infection



# 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
- ▶  $\delta$  : importation rate (which we assume is 0)



# Champagne Model Transition Rates

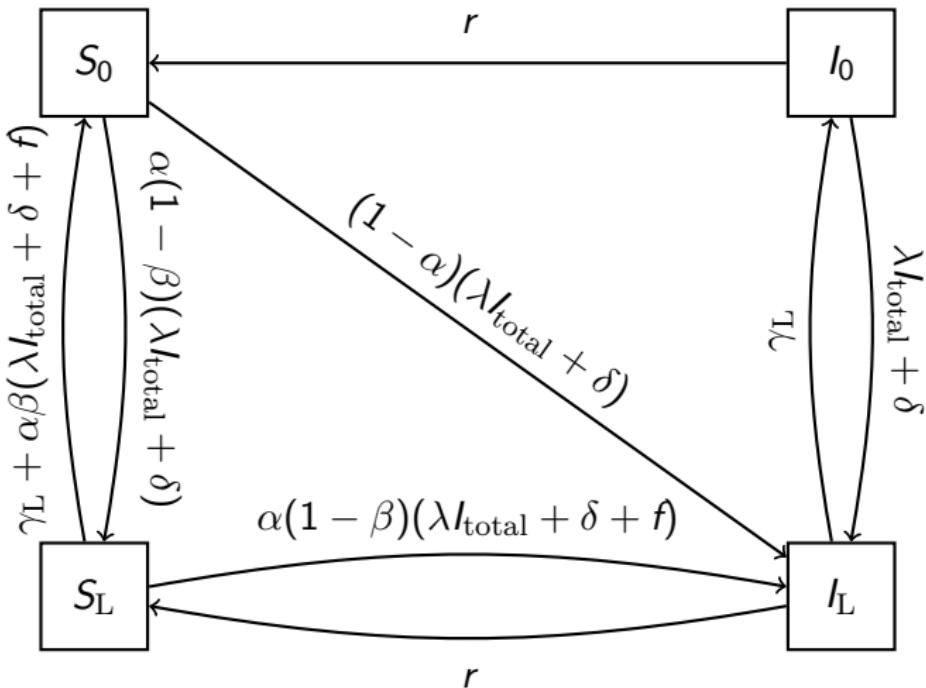


Figure: *P. vivax* model described by Champagne et al. 2022

# Champagne ODEs

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

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

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

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



# Model Calibration Data

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



# Model Calibration Data

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

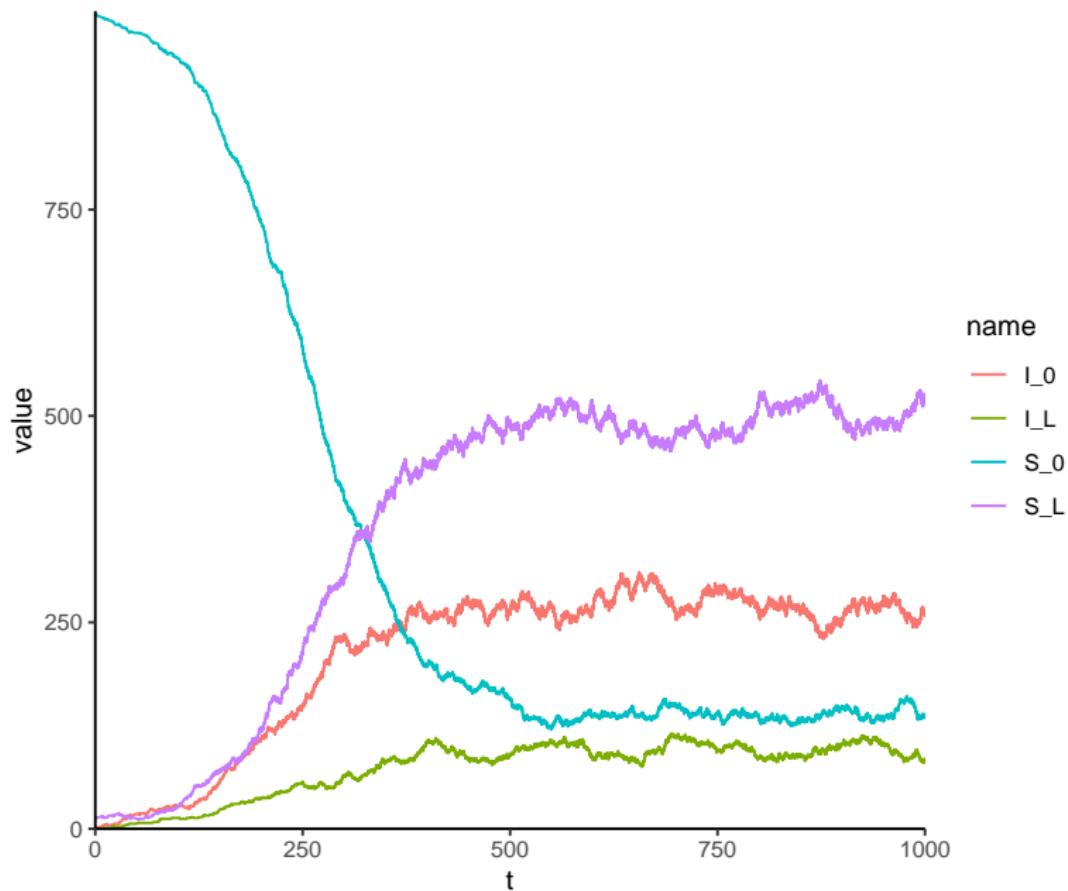


# Model Calibration Data

- ▶ Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.
- ▶  $S(\mathbf{X}_{\text{obs}}) := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$ 
  - ▶  $w_{\text{obs}}$  : weekly incidence around (stochastic) equilibrium
  - ▶  $p_{\text{obs}}$  : prevalence around (stochastic) equilibrium
  - ▶  $m_{\text{obs}}$  : incidence in the first month of the epidemic
- ▶  $\mathcal{D}(\boldsymbol{\theta}) := \left| \frac{w_{\text{obs}} - w}{w_{\text{obs}}} \right| + \left| \frac{p_{\text{obs}} - p}{p_{\text{obs}}} \right| + \left| \frac{m_{\text{obs}} - m}{m_{\text{obs}}} \right|$ 
  - ▶ ( $L_1$  norm on the relative differences)



# Example Simulation



# What's the Bayesian part?

- ▶ Bayesian acquisition function  $\arg \min_{\theta} A(\theta)$



# What's the Bayesian part?

- ▶ Bayesian acquisition function  $\arg \min_{\theta} A(\theta)$
- ▶ Gutmann and Cor 2016 uses

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

- ▶  $\eta_t := \sqrt{c + 2 \ln(t^{d/2+2})}$ , and  $c$  can be chosen
- ▶  $\mu(\theta)$  and  $v(\theta)$  are the posterior mean and variance



# What's the Bayesian part?

- ▶ Bayesian acquisition function  $\arg \min_{\theta} A(\theta)$
- ▶ Gutmann and Cor 2016 uses

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

- ▶  $\eta_t := \sqrt{c + 2 \ln(t^{d/2+2})}$ , and  $c$  can be chosen
- ▶  $\mu(\theta)$  and  $v(\theta)$  are the posterior mean and variance
- ▶ Could use expected information

$$(\mu_{\min} - \mu(\theta))\Phi\left(\frac{\mu_{\min} - \mu(\theta)}{\sqrt{v(\theta)}}\right) + \sqrt{v(\theta)}\phi\left(\frac{\mu_{\min} - \mu(\theta)}{\sqrt{v(\theta)}}\right)$$

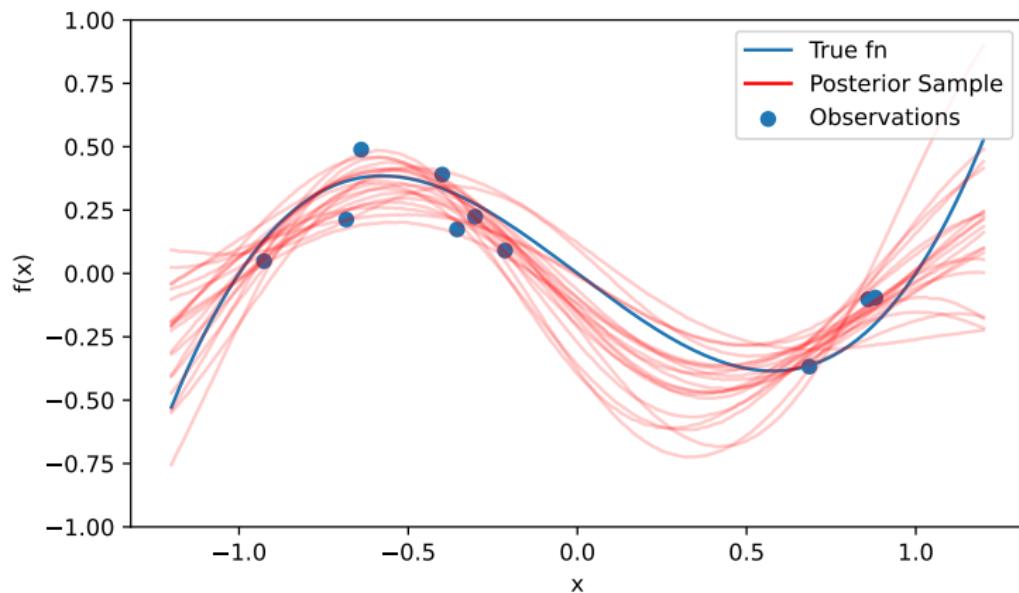
- ▶  $\mu_{\min} := \min_{\theta} \mu(\theta)$
- ▶  $\Phi, \phi$  CDF and PDF of standard normal



# Acquisition 'example'



## Acquisition ‘example’



# Synthetic Likelihood

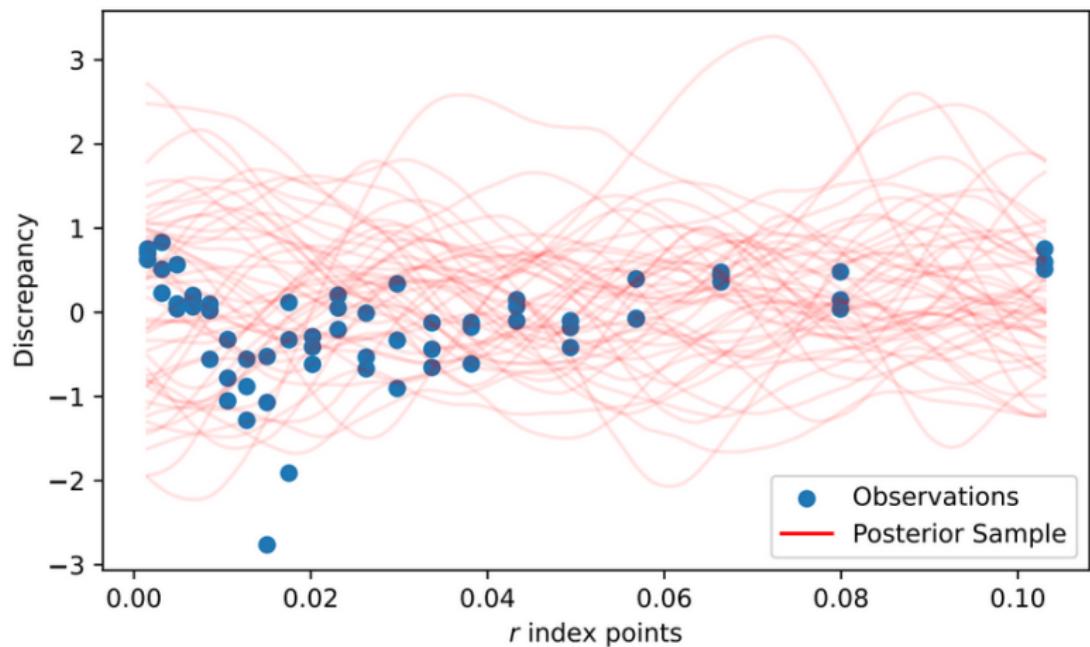
- ▶  $L(\theta | \mathbf{X}_{\text{obs}}) \approx P(\mathcal{D}_{\mathcal{GP}}(\theta) < \varepsilon)$  (up to a proportion) where  $\mathcal{D}_{\mathcal{GP}}$  is the discrepancy modelled the Gaussian process



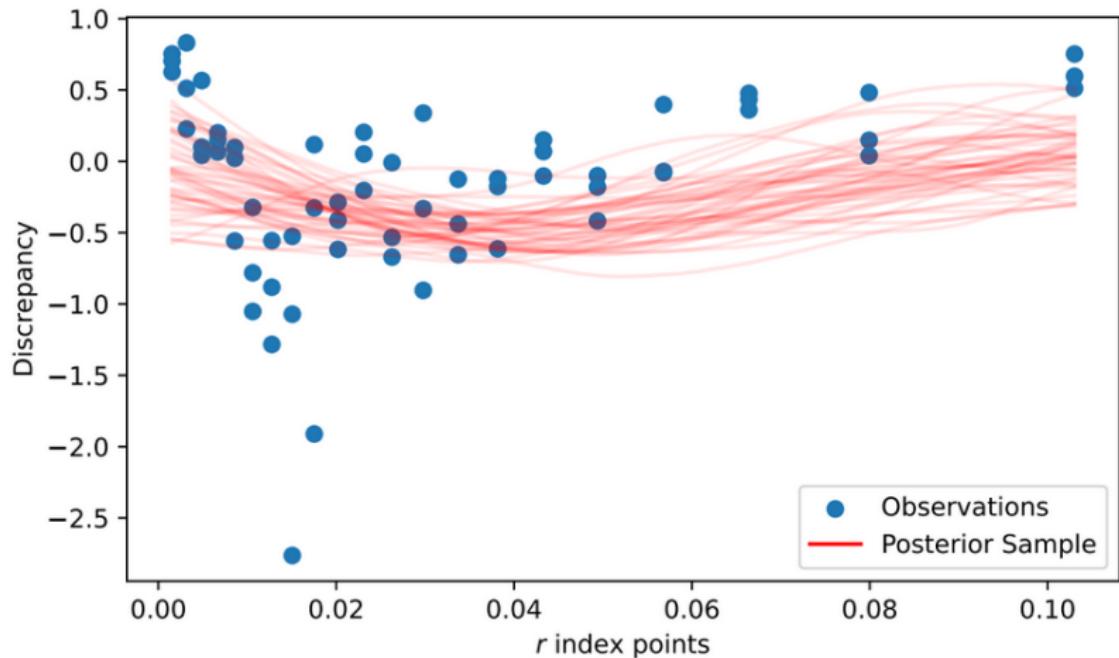
# How did it go?



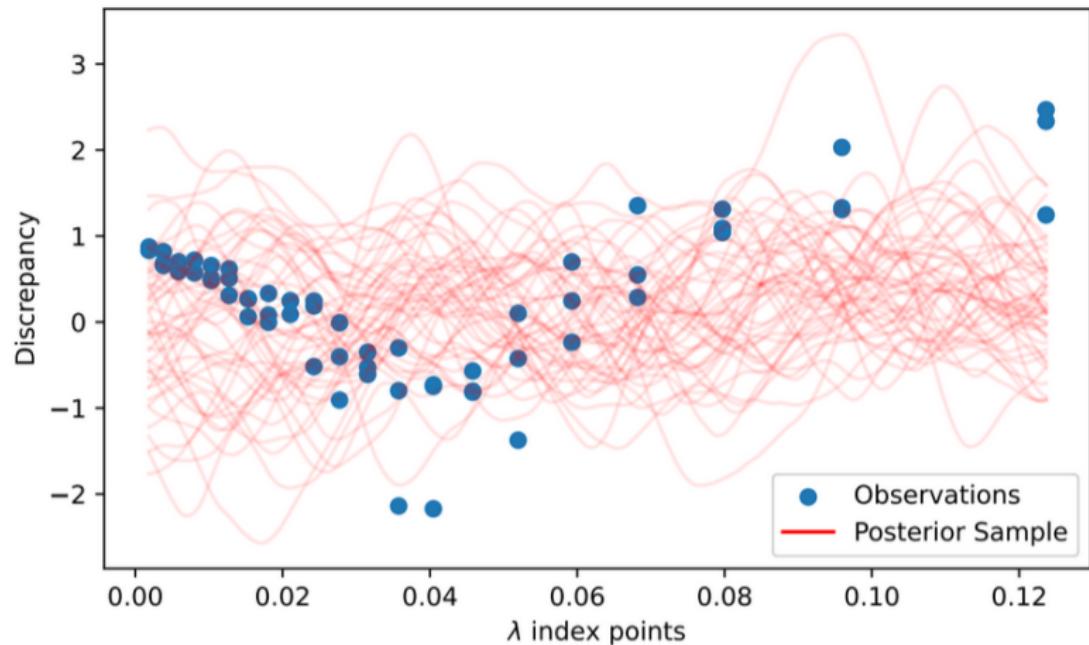
# How did it go?



# How did it go?



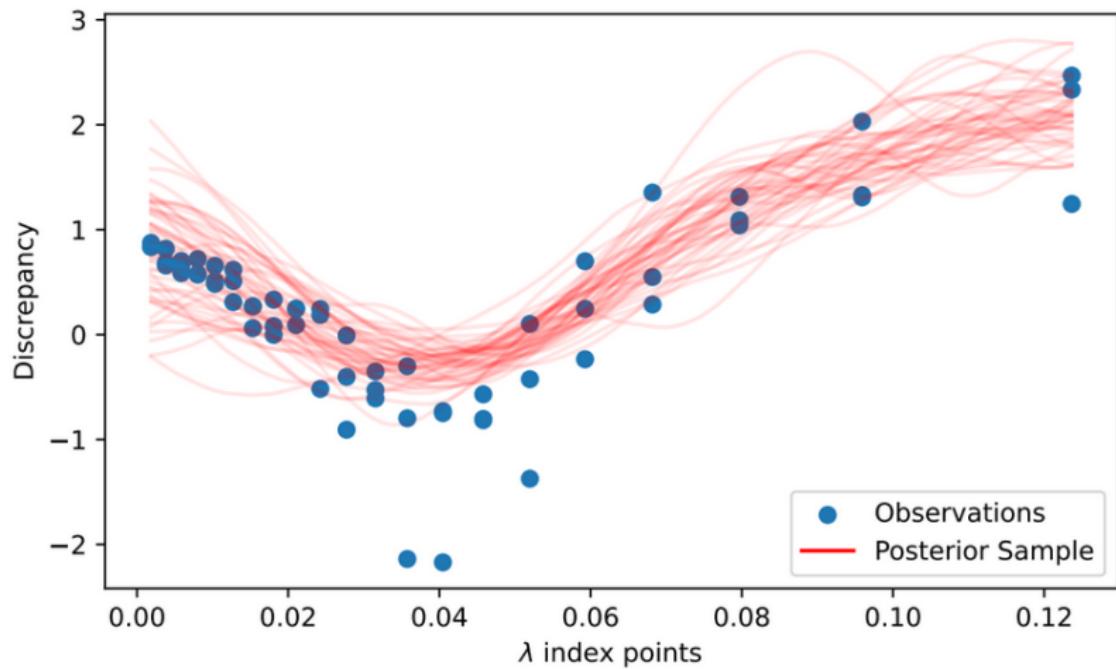
# How did it go?



THE UNIVERSITY OF  
MELBOURNE



# How did it go?



# Big Problem (Big Solutions?)

- ▶ Observation variance is considered constant across the GP
  - ▶ Particularly a problem at the threshold
- ▶ Assumes that normal distribution approximates  $\mathcal{D}(\theta)$ 
  - ▶ Log-normal might be more realistic?
- ▶ Jumps where there is threshold/bifurcation behaviour
  - ▶ Use Student  $t$ -Process to allow for jumps?



# Big Problem (Big Solutions?)

- ▶ Observation variance is considered constant across the GP
  - ▶ Particularly a problem at the threshold
  - ▶ Fix by modelling observation variance as another GP
- ▶ Assumes that normal distribution approximates  $\mathcal{D}(\theta)$ 
  - ▶ Log-normal might be more realistic?
- ▶ Jumps where there is threshold/bifurcation behaviour
  - ▶ Use Student  $t$ -Process to allow for jumps?



# Thanks to

- ▶ Eamon Conway
- ▶ Jennifer Flegg
- ▶ You, since I didn't have to explain what Vivax malaria was :)

