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 is considered a (random) function $f(\boldsymbol{\theta})$ that maps $\boldsymbol{\theta}$ (a vector of parameters) to a model output, that can be transformed into $\mathbf{X}$, that has the same shape as:
- $\mathbf{X}_{\text{obs}}$, a vector of outputs given to us usually in the forms of summary statistics (incidence, prevalence, hospitalisations etc).

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 even $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$, where $S(\mathbf{X}_{\text{obs}})$ is a (vector of) summary statistic(s)

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 even $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$, where $S(\mathbf{X}_{\text{obs}})$ is a (vector of) summary statistic(s)
- $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}}))$

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 even $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$, where $S(\mathbf{X}_{\text{obs}})$ is a (vector of) summary statistic(s)
- $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}}))$
- $\Pr(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) \propto \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$

The Sad Truth

- As models become more complicated, explicit likelihoods don't exist (think agent based models).

A Standard Bayesian Solution

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

What is 'matches'

- Discrepency function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  - Can be a norm such as $||S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$

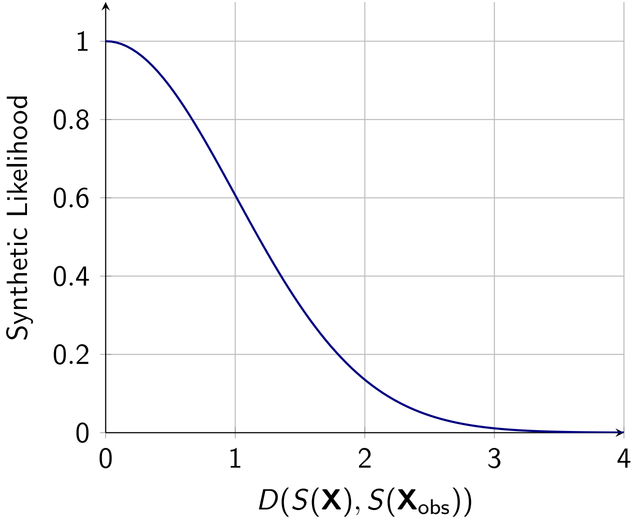
What is 'matches'

- Discrepency function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  - Can be a norm such as $||S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$
  - Care should be taken to rescale $S(\mathbf{X}_{\text{obs}})$ and $S(\mathbf{X})$ appropriately (ie via a covariance matrix).

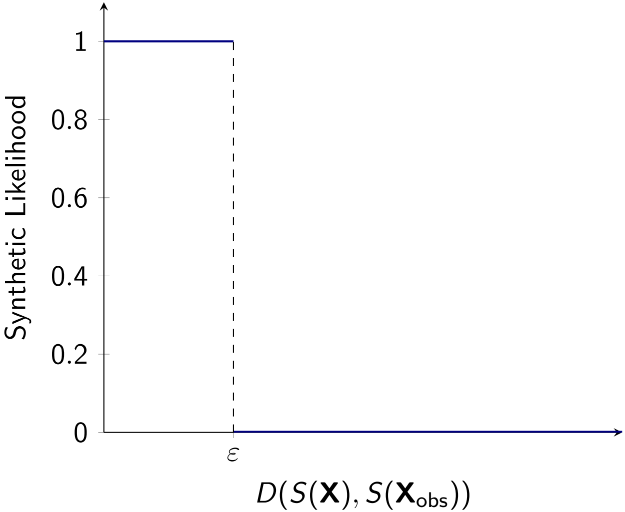
What is 'matches'

- Discrepency function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  - Can be a norm such as $||S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$
  - Care should be taken to rescale $S(\mathbf{X}_{\text{obs}})$ and $S(\mathbf{X})$ appropriately (ie via a covariance matrix).
- $D(S(\mathbf{X}), S(\mathbf{X}_{\text{obs}}))$, gives acceptance probability of $\boldsymbol{\theta}$.

Acceptance Probability



Uniform Acceptance Probability



Overall Idea of my Research

- What if we could 'predict' discrepancy values we hadn't seen before?

Overall Idea of my Research

- What if we could 'predict' discrepency values we hadn't seen before?
- For parameters 'close' to parameters we've already tried it should be easy.

Gaussian Processes

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

Gaussian Processes on $\mathbb{R}^d$

Definition (Gaussian Process)

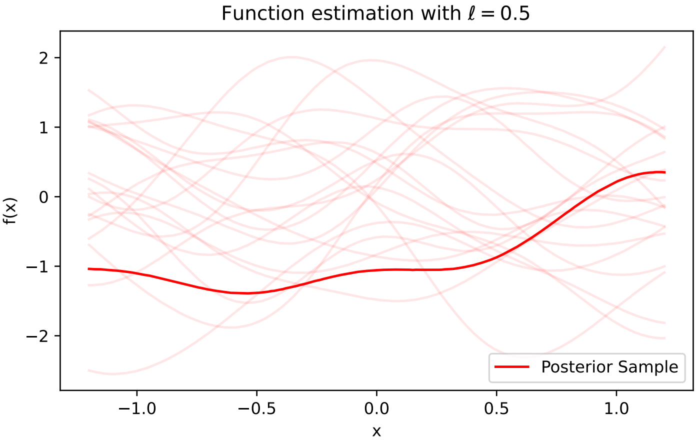
A collection of random variables $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ *is a Gaussian process if all finite dimensional distributions are multivariate normal distributed. That is, there is a function* $m : \mathbf{x} \to \mathbb{R}$ *and kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that for all finite sets* $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$,

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

where

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

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

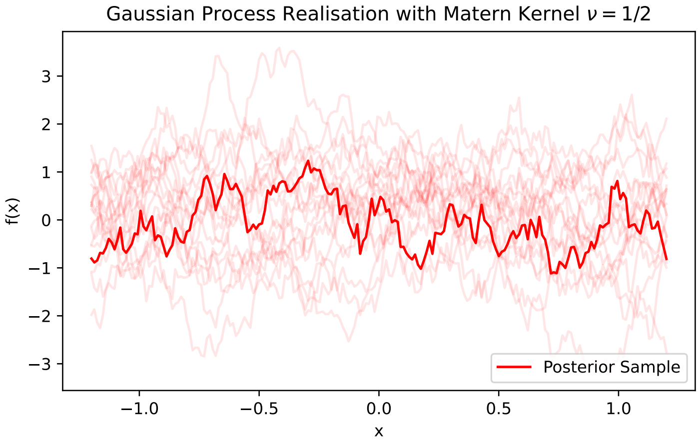
- Matern Kernel

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

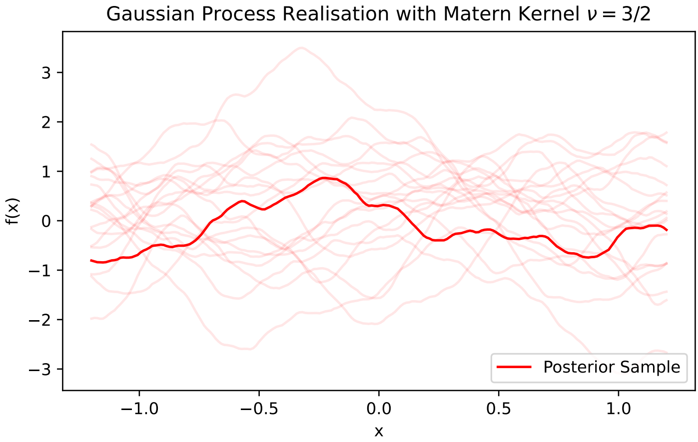
  where K_ν is a modified Bessel function ($|| \cdot ||$ is the euclidean distance)
- $\lfloor \nu \rfloor$ times mean square differentiable.
- As $\nu \to \infty$ you get squared exponential covariance function, which results in realisations that are infinitely mean square differentiable:

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

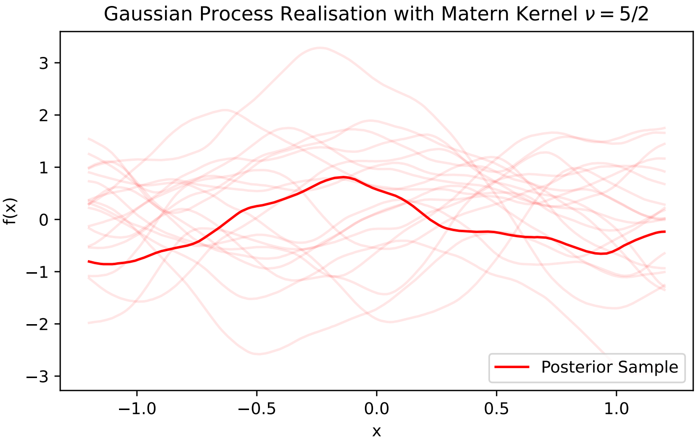
Kernel Choices - Kernel Type



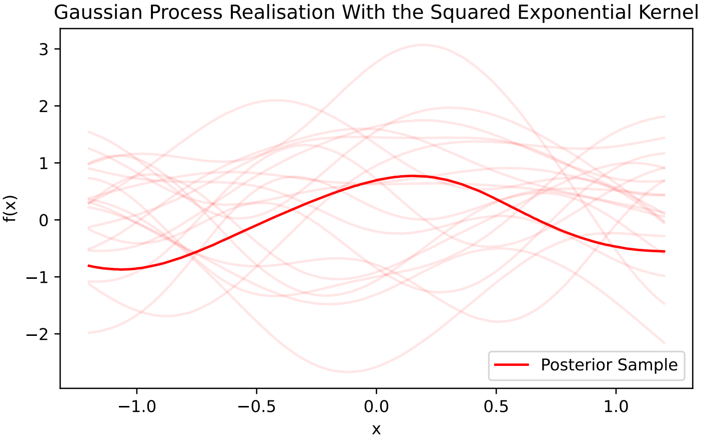
Kernel Choices - Kernel Type



Kernel Choices - Kernel Type



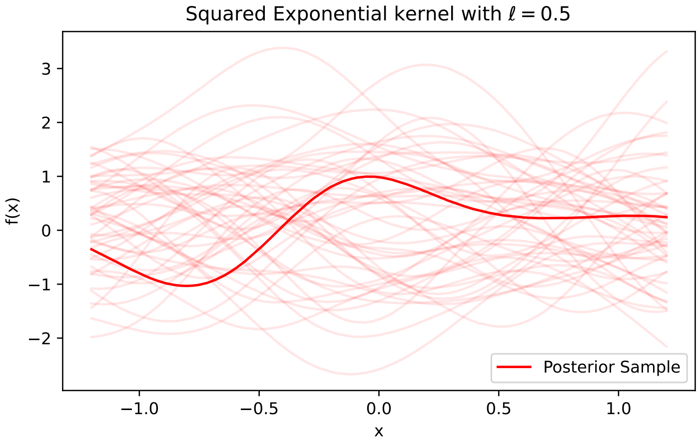
Kernel Choices - Kernel Type



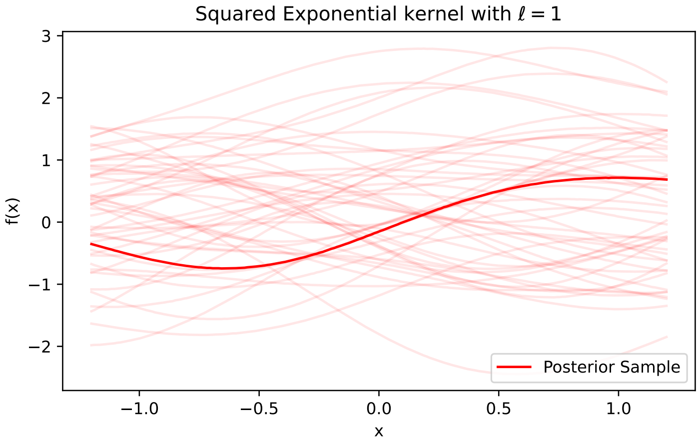
Kernel Choices - length scale

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

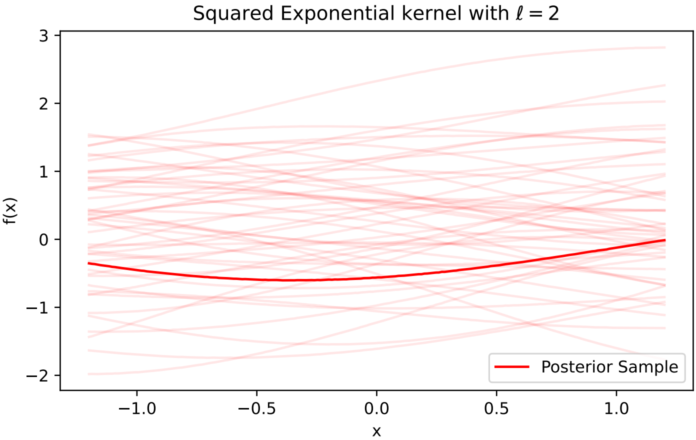
Kernel Choices - length scale



Kernel Choices - length scale



Kernel Choices - length scale

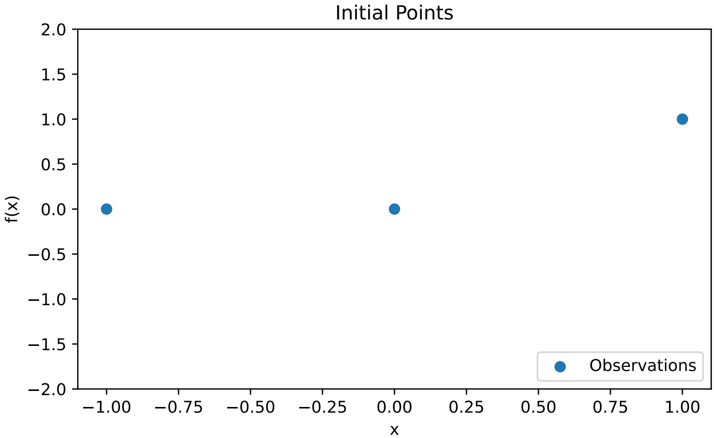


Fitting our GP to data

GPs are 'priors'

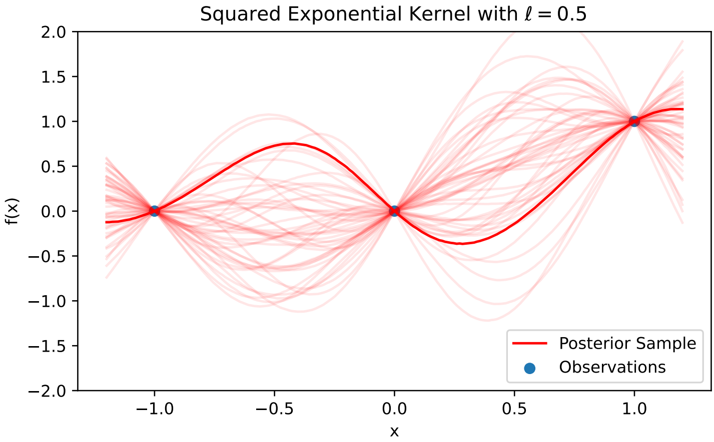
Fitting our GP to data

GPs are 'priors'



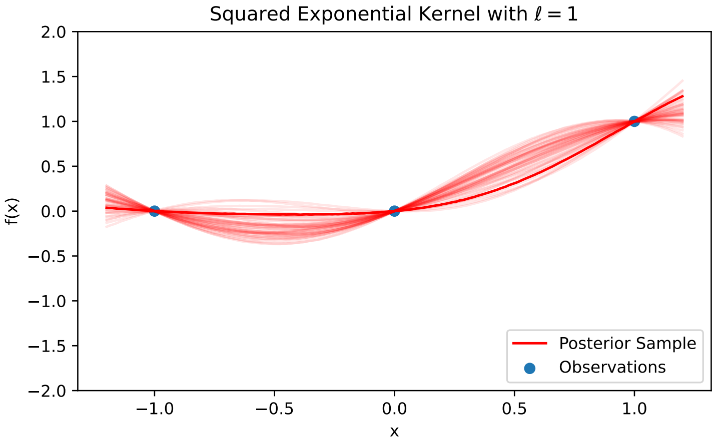
Fitting our GP to data

GPs are 'priors'



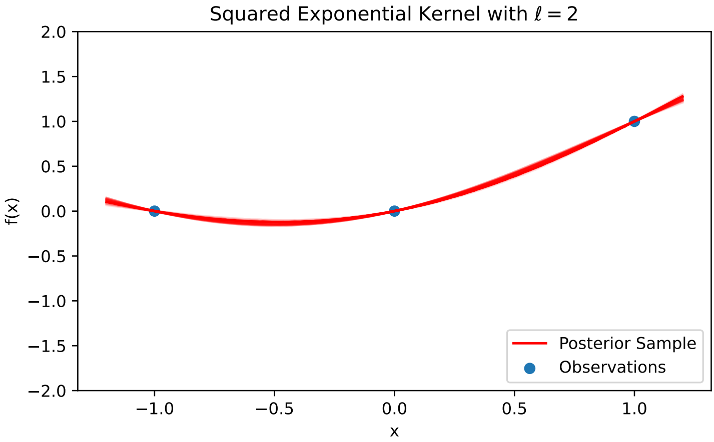
Fitting our GP to data

GPs are 'priors'



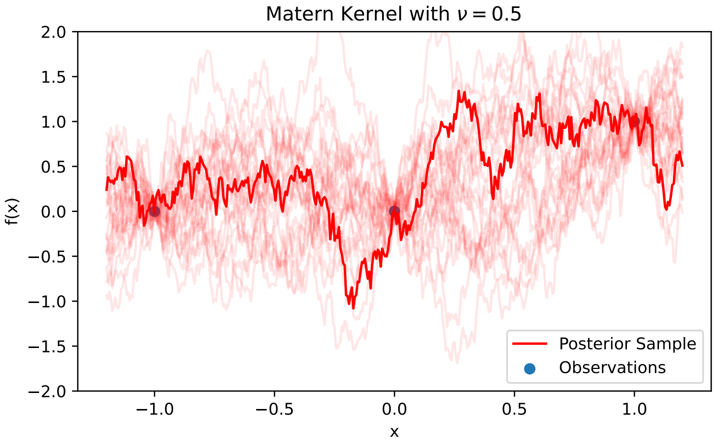
Fitting our GP to data

GPs are 'priors'



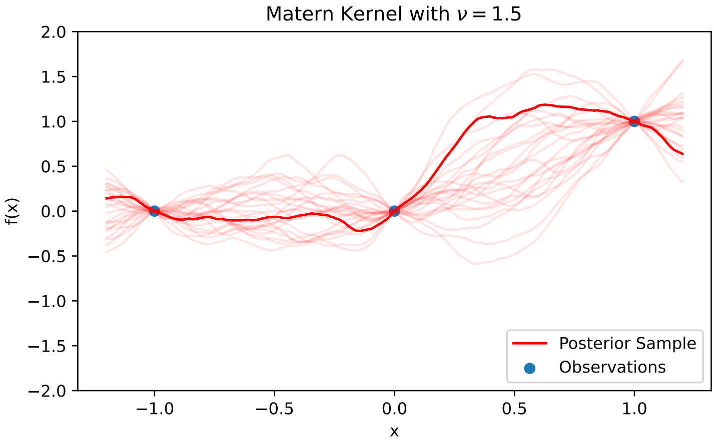
Fitting our GP to data

GPs are 'priors'



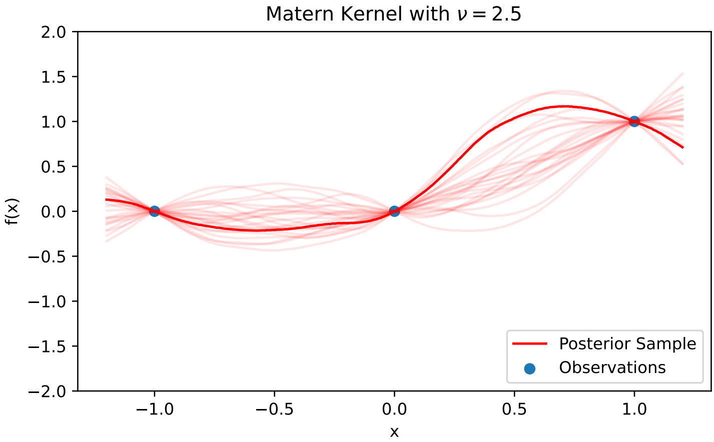
Fitting our GP to data

GPs are 'priors'



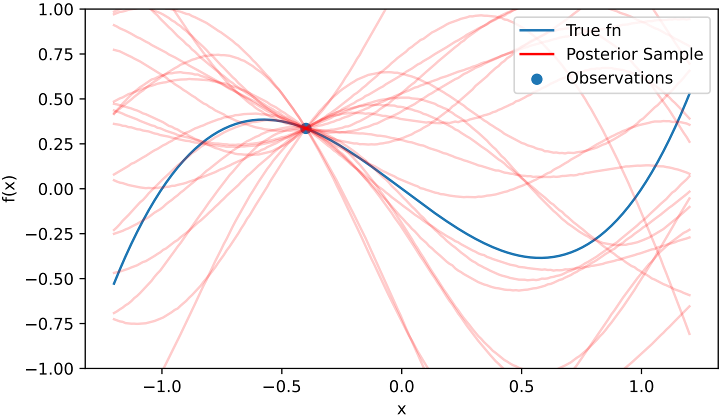
Fitting our GP to data

GPs are 'priors'

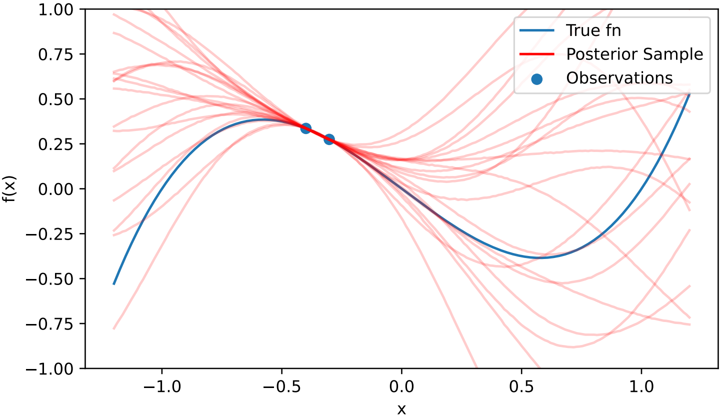


GP regression on $x(x-1)(x+1)$

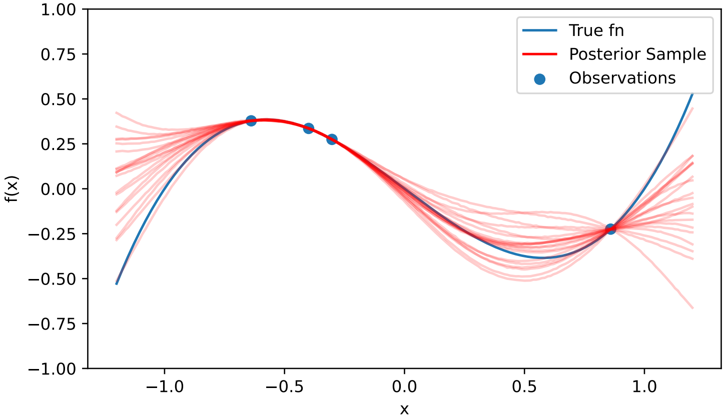
GP regression on $x(x-1)(x+1)$



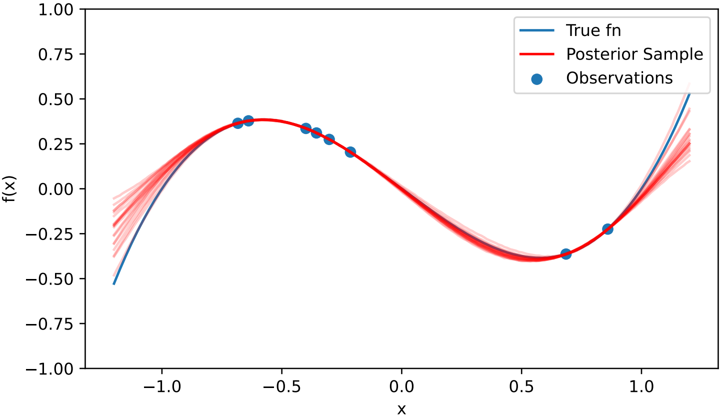
GP regression on $x(x-1)(x+1)$



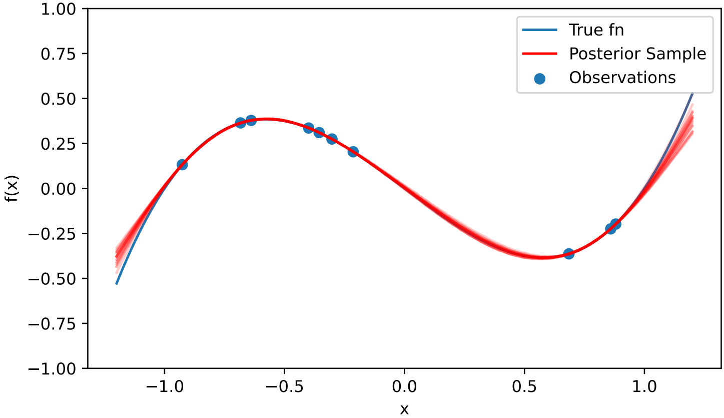
GP regression on $x(x-1)(x+1)$



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



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



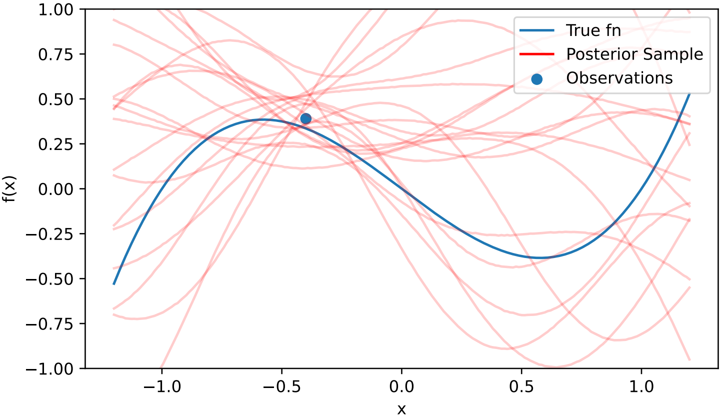
What if we have noise?

Add in observation variance, so that

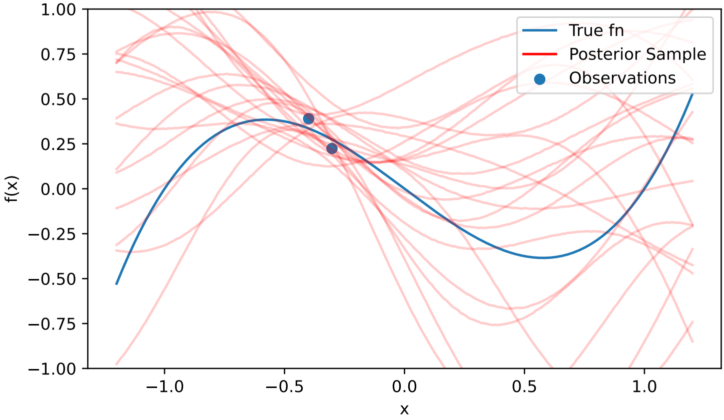
$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} + \sigma^2 \mathbf{I}_n \right)$$

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

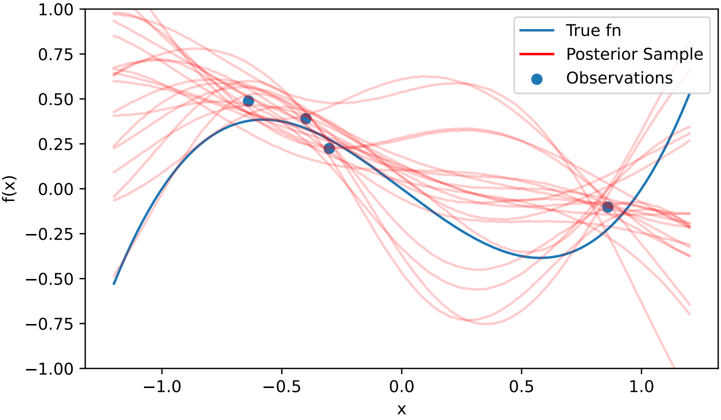
GP regression on $x(x-1)(x+1)$



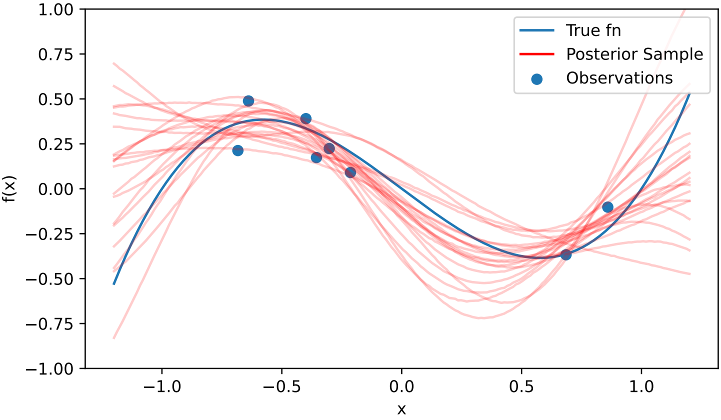
GP regression on $x(x-1)(x+1)$



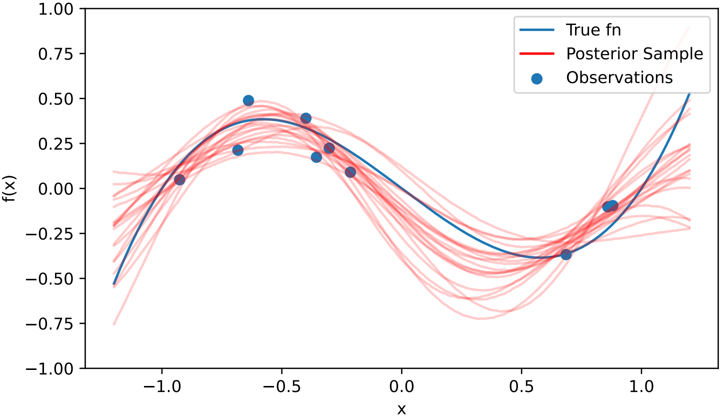
GP regression on $x(x-1)(x+1)$



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



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



Overall Idea again

- What if we could 'predict' discrepency values we hadn't seen before?

Overall Idea again

- What if we could 'predict' discrepency values we hadn't seen before?
- For parameters 'close' to parameters we've already tried it should be easy

Overall Idea again

- What if we could 'predict' discrepency values we hadn't seen before?
- Use Gaussian process to predict discrepency function

About Vivax Malaria

- Has dormant liver stage on top of blood stage infection

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
- δ : importation rate (which we assume is 0)

Champagne ODEs

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

Champagne Model Diagram

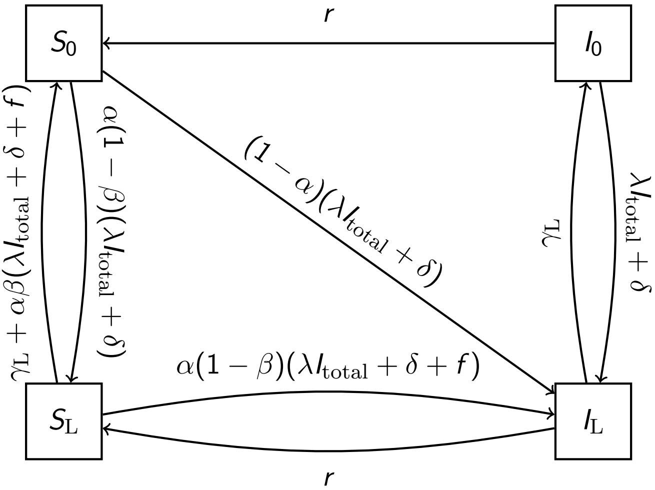


Figure: *P. vivax* model described by Champagne et al. 2022

Model Calibration Data

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

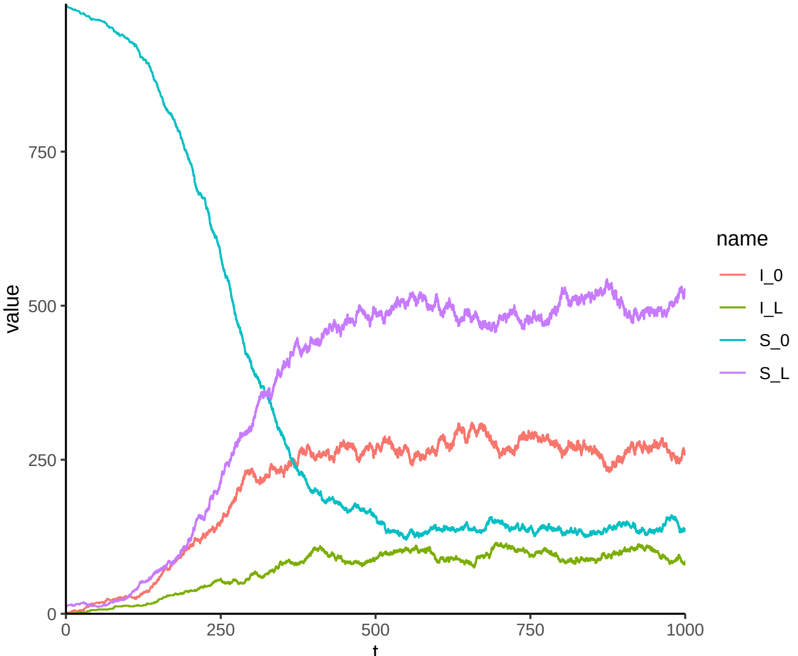
Model Calibration Data

- 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_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{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_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{obs} : incidence in the first month of the epidemic
- $D(S(\mathbf{X}), S(\mathbf{X}_{\text{obs}})) := \left|\frac{w_{\text{obs}} - w}{w_{\text{obs}}}\right| + \left|\frac{p_{\text{obs}} - p}{p_{\text{obs}}}\right| + \left|\frac{m_{\text{obs}} - m}{m_{\text{obs}}}\right|$
  - L_1 norm on the relative differences

Example Simulation



What's the Bayesian part?

- Choosing the next point to sample

What's the Bayesian part?

- Choosing the next point to sample
- $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$ where A is an acquisition function.

What's the Bayesian part?

- Choosing the next point to sample
- $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$ where A is an acquisition function.
- BOLFI paper uses

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

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

What's the Bayesian part?

- Choosing the next point to sample
- $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$ where A is an acquisition function.
- BOLFI paper uses

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

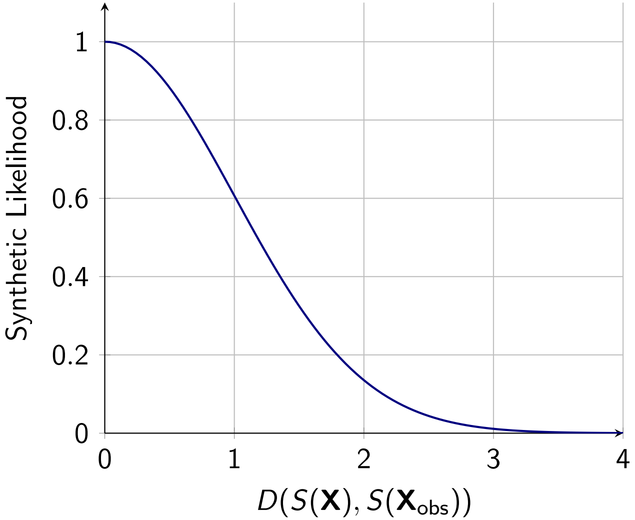
  - $\eta_t := \sqrt{c + 2\ln(t^{d/2+2})}$, and c can be chosen
  - $\mu(\boldsymbol{\theta})$ and $\mathrm{v}(\boldsymbol{\theta})$ are the posterior mean and variance
- Could use expected information

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

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

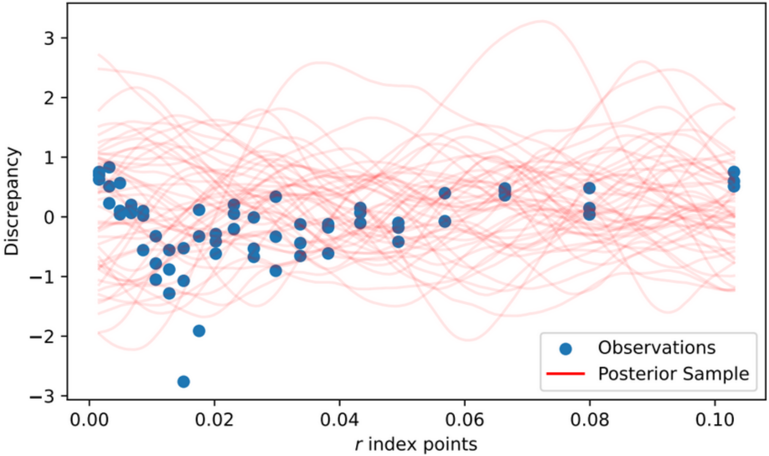
Synthetic Likelihood

- $L(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) \approx P(D_{\mathcal{GP}}(\theta) = 0)$ where $D_{\mathcal{GP}}$ is the discrepency modelled the Gaussian process
- This is equivalent to using the half normal acceptance criteria



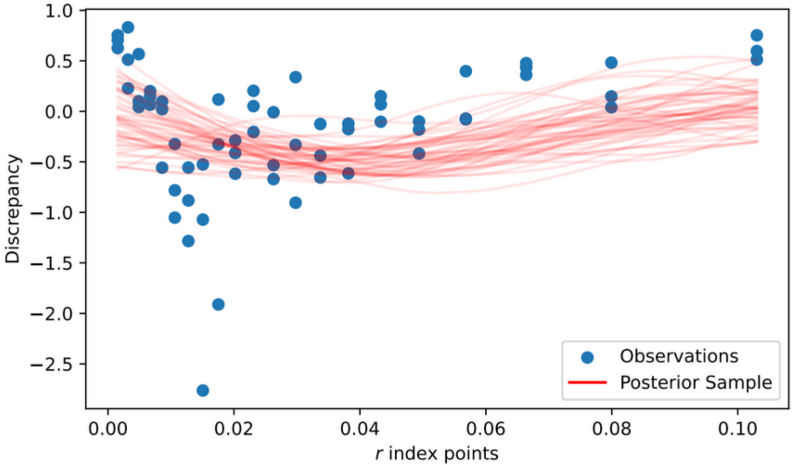
How did it go?

How did it go?



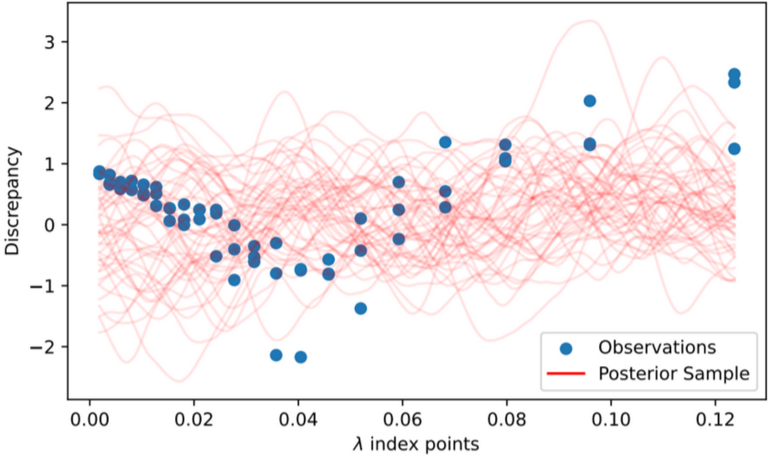
Discrepancy
r index points
0.00
0.02
0.04
0.06
0.08
0.10
3
2
1
0
−1
−2
−3
Observations
Posterior Sample

How did it go?



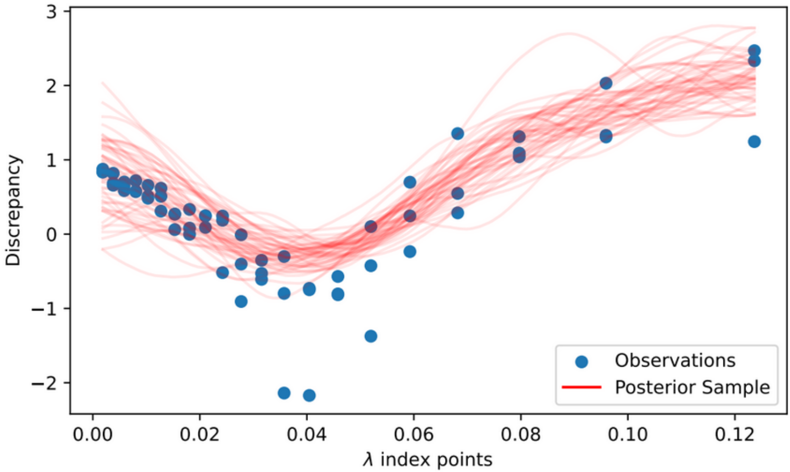
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
Discrepancy
0.00
0.02
0.04
0.06
0.08
0.10
r index points
Observations
Posterior Sample

How did it go?



3
2
1
0
−1
−2
Discrepancy
0.00
0.02
0.04
0.06
0.08
0.10
0.12
λ index points
Observations
Posterior Sample

How did it go?



Observations
Posterior Sample
Discrepancy
λ index points
3
2
1
0
−1
−2
0.00
0.02
0.04
0.06
0.08
0.10
0.12

Big Problem (Big Solutions?)

- Observation variance is considered constant across the GP
  - Particularly a problem at the threshold

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

Thanks to

- Eamon Conway and the Mueller lab at WEHI
- Jennifer Flegg at Unimelb
- Damon for explaining disease modelling so I don't have to