Bayesian Optimisation for Likelihood Free Inference

Make model parameterisation go brrr

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

May 2024

Notation

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

Notation

- Model a (random) function $f\colon \boldsymbol{\Theta} \to \boldsymbol{\mathcal{Y}}$.
  - $\boldsymbol{\Theta}$: parameter space
  - $\boldsymbol{\mathcal{Y}}$: model output space
  - $\mathbf{Y}_{\boldsymbol{\theta}} := f(\boldsymbol{\theta})$ (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 $f : \boldsymbol{\Theta} \to \boldsymbol{\mathcal{Y}}$.
  - $\boldsymbol{\Theta}$: parameter space
  - $\boldsymbol{\mathcal{Y}}$: model output space
  - $\mathbf{Y}_{\boldsymbol{\theta}} := f(\boldsymbol{\theta})$ (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 $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{Y}_{\text{obs}})) := \Pr(S(\mathbf{Y}_{\text{obs}})|\boldsymbol{\theta})$

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

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

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

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

THE UNIVERSITY OF MELBOURNE
WEHI
brighter together

Reality

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

A Standard Bayesian Solution

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

What is 'matches'

- Option 1: $\mathbf{Y}_{\boldsymbol{\theta}_i} = \mathbf{Y}_{\text{obs}}$ or $S(\mathbf{Y}_{\boldsymbol{\theta}_i}) = S(\mathbf{Y}_{\text{obs}})$

What is 'matches'

- Option 1: $\mathbf{Y}_{\boldsymbol{\theta}_i} = \mathbf{Y}_{\text{obs}}$ or $S(\mathbf{Y}_{\boldsymbol{\theta}_i}) = S(\mathbf{Y}_{\text{obs}})$
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - e.g. p-norm

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

What is 'matches'

- Option 1: $\mathbf{Y}_{\boldsymbol{\theta}_i} = \mathbf{Y}_{\text{obs}}$ or $S(\mathbf{Y}_{\boldsymbol{\theta}_i}) = S(\mathbf{Y}_{\text{obs}})$
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - e.g. p-norm

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

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

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

What is 'matches'

- Option 1: $\mathbf{Y}_{\boldsymbol{\theta}_i} = \mathbf{Y}_{\text{obs}}$ or $S(\mathbf{Y}_{\boldsymbol{\theta}_i}) = S(\mathbf{Y}_{\text{obs}})$
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - e.g. p-norm

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

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

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

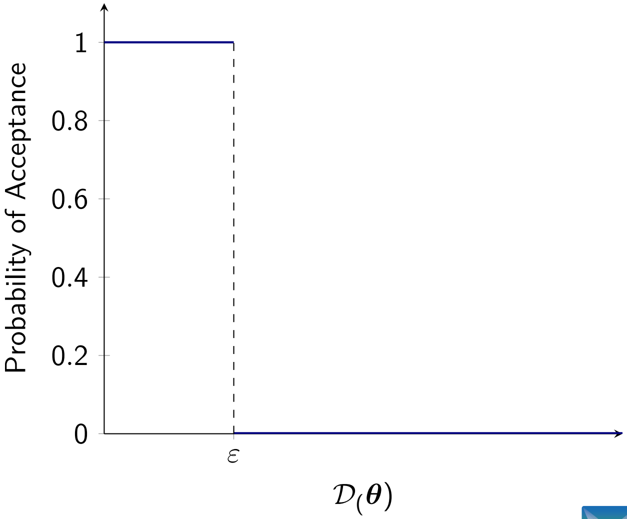
What is 'matches'

- Option 1: $\mathbf{Y}_{\boldsymbol{\theta}_i} = \mathbf{Y}_{\text{obs}}$ or $S(\mathbf{Y}_{\boldsymbol{\theta}_i}) = S(\mathbf{Y}_{\text{obs}})$
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - e.g. p-norm

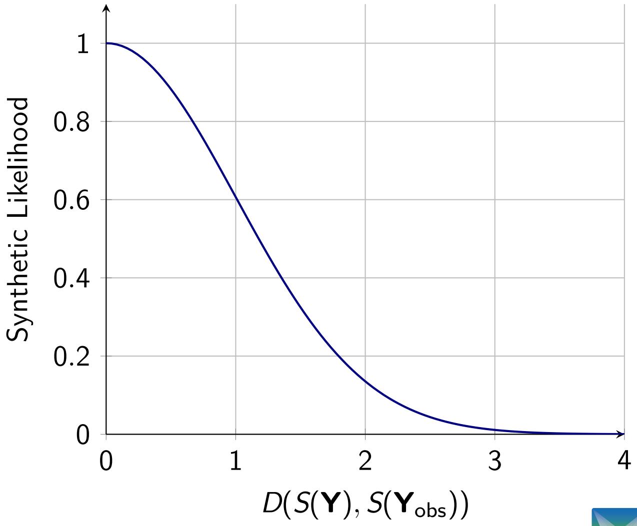
$$||S(\mathbf{Y}_{\boldsymbol{\theta}_i}) - S(\mathbf{Y}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{Y}_{\boldsymbol{\theta}_i}) - S(\mathbf{Y}_{\text{obs}})|^p)^{1/p}$$

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

Uniform Acceptance Probability



Acceptance Probability



Overall Idea of my Research

- Can we predict $\mathcal{D}(\boldsymbol{\theta}_i)$ without having to evaluate $f(\boldsymbol{\theta}_i)$?

Overall Idea of my Research

- Can we predict $\mathcal{D}(\boldsymbol{\theta}_i)$ without having to evaluate $f(\boldsymbol{\theta}_i)$?
- 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* $m : \mathbf{x} \to \mathbb{R}$ *and kernel* $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *such that for all finite sets* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

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

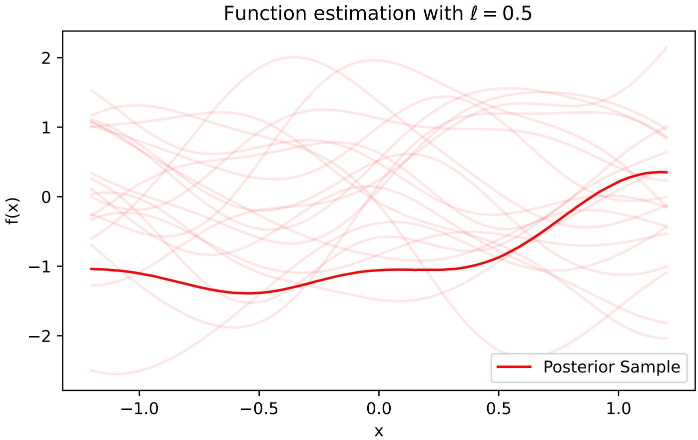
where

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

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

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_ν is a modified Bessel function ($||\cdot||$ is the euclidean distance)
- $\lfloor \nu \rfloor$ times mean square differentiable.
- $\nu \to \infty$ - infinitely mean square differentiable squared exponential covariance kernel (strong assumption)

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

Kernel Classes

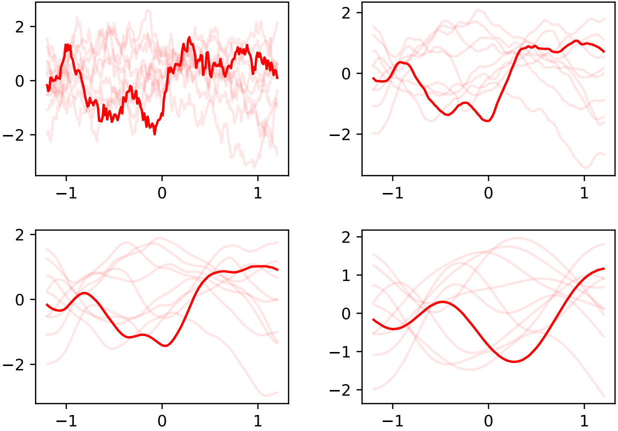


Figure: Matérn 1/2, 3/2, 5/2, and squared exponential kernels.

Gaussian Process Regression

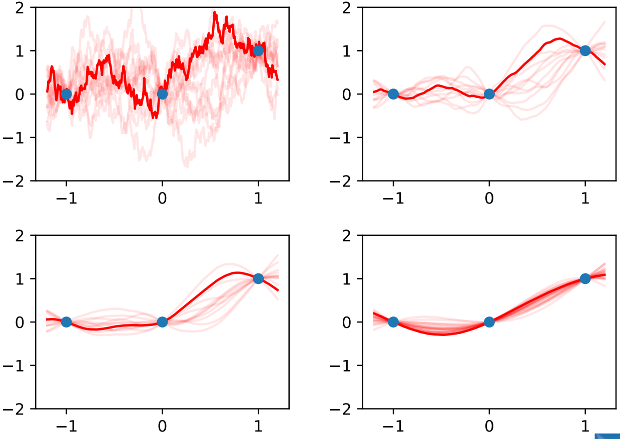
$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}\right)$$

implies

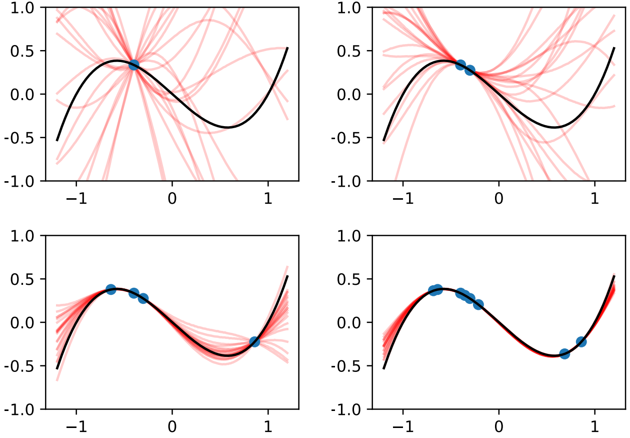
$$f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_* K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\; K - K_* K_{**}^{-1} K_*^T\right).$$

Fitting our GP to data

GPs are 'priors'



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

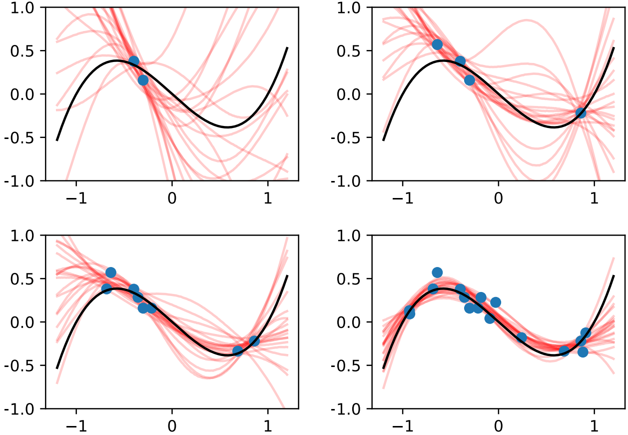


What if we have noise?

Add observation variance σ_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 $\mathcal{D}(\boldsymbol{\theta}_i)$ without having to evaluate $f(\boldsymbol{\theta}_i)$ <2->
- $\mathcal{D}(\boldsymbol{\theta}) \approx \mathcal{D}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ close. <3->
- Approximate $\mathcal{D}(\boldsymbol{\theta})$ by a Gaussian process $\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta})$

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Bayesian Acquisition

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

Bayesian Acquisition

- Gutmann and Cor 2016 uses lower confidence bound

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

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

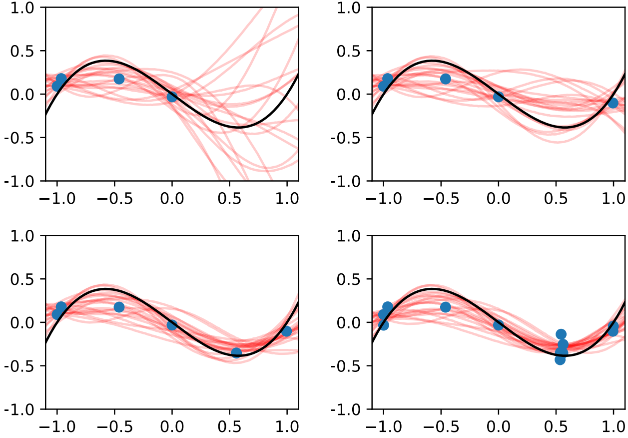
$$
\begin{aligned}
A_{\mathsf{EI}}(\boldsymbol{\theta}) :=& \mathbb{E}(\min[\mathcal{D}_{\mathcal{GP}}(\boldsymbol{\theta}) - \mu_{\mathsf{min}}, 0)] \\
=& (\mu_{\mathsf{min}} - \mu(\boldsymbol{\theta}))\Phi\left(\frac{\mu_{\mathsf{min}} - \mu(\boldsymbol{\theta})}{\sqrt{\mathrm{v}(\boldsymbol{\theta})}}\right)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

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

Bayesian Acquisition

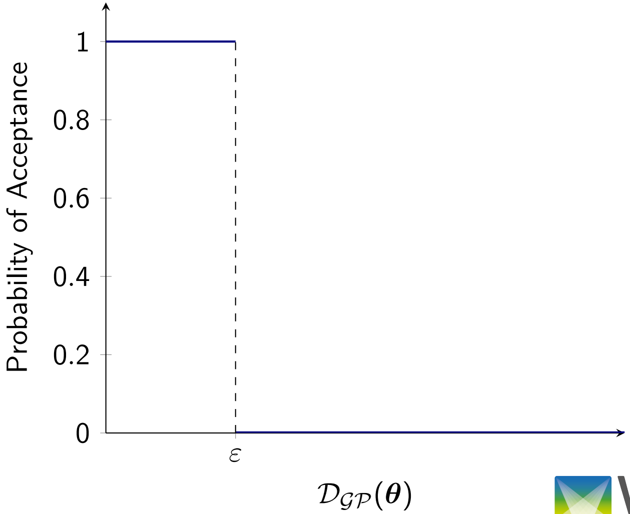
- Theoretical guarantees highly sensitive to choice of kernel

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

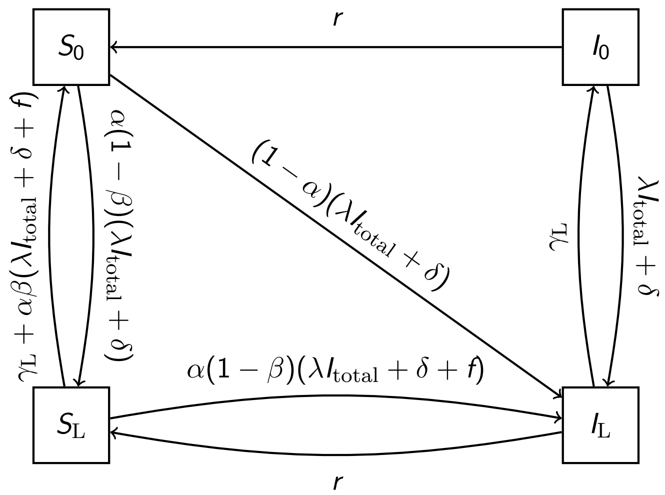


Figure: *P. vivax* model described by Champagne et al. 2022

Ordinary Differential Equations - Champagne et. al

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

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

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)

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_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{obs} : incidence in the first month of the epidemic

- Expected information acquisition function

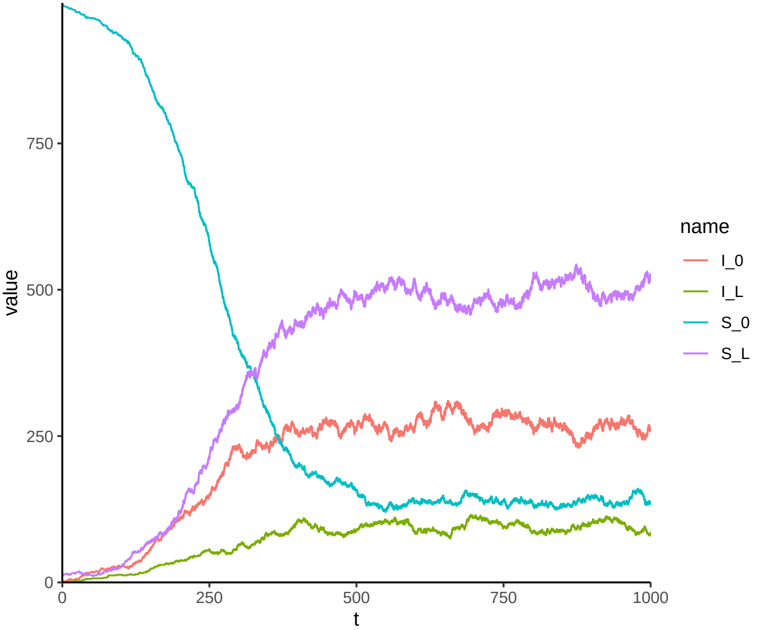
Model Calibration Data

- 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_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{obs} : incidence in the first month of the epidemic
-

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

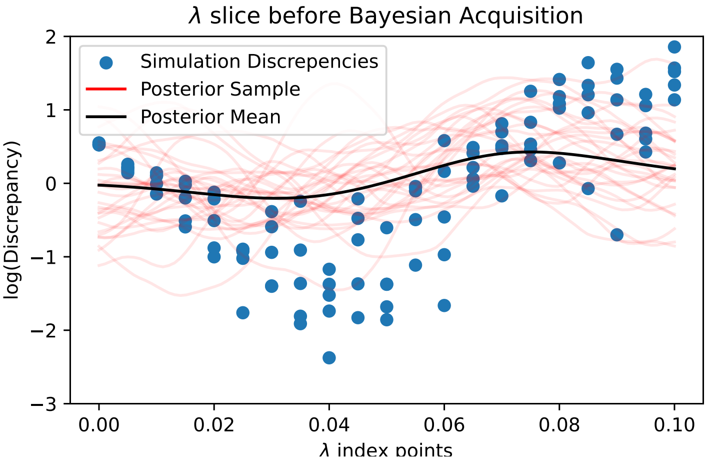
  - (Log of the L_2 norm of the relative differences)
- Expected information acquisition function

Example Simulation



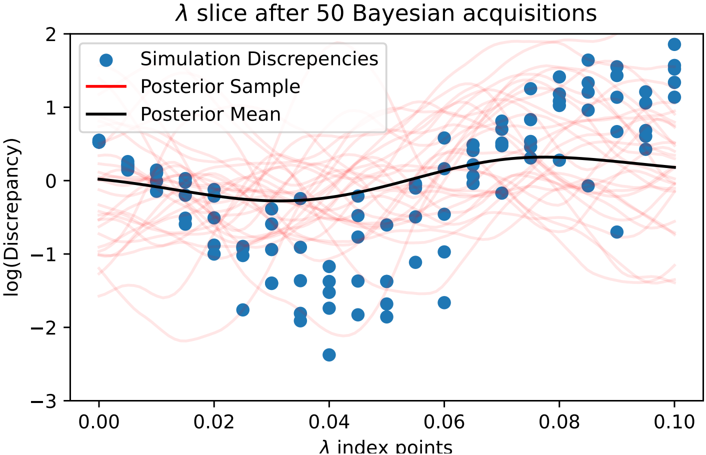
How did it go?

How did it go?



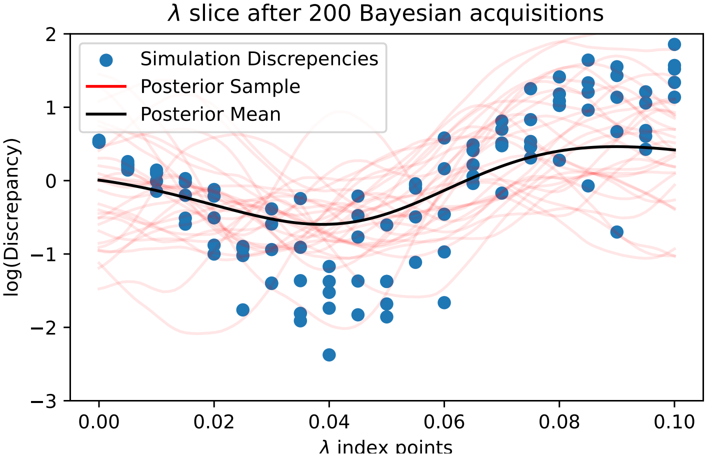
λ slice before Bayesian Acquisition
Simulation Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
2
1
0
−1
−2
−3
0.00
0.02
0.04
0.06
0.08
0.10
λ index points

How did it go?



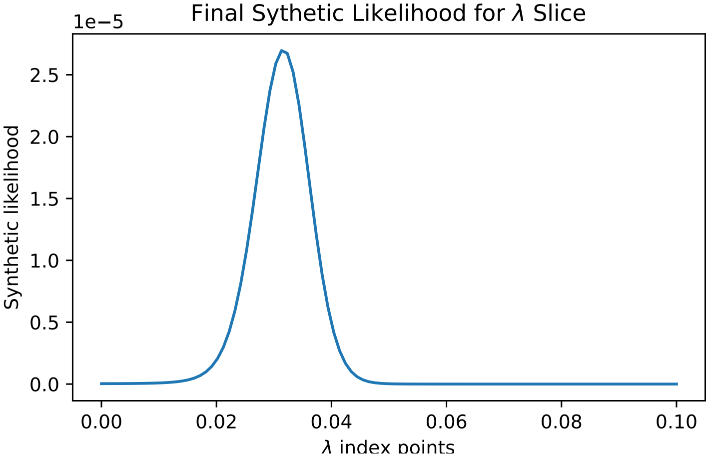
λ slice after 50 Bayesian acquisitions
Simulation Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
2
1
0
−1
−2
−3
0.00
0.02
0.04
0.06
0.08
0.10
λ index points

How did it go?

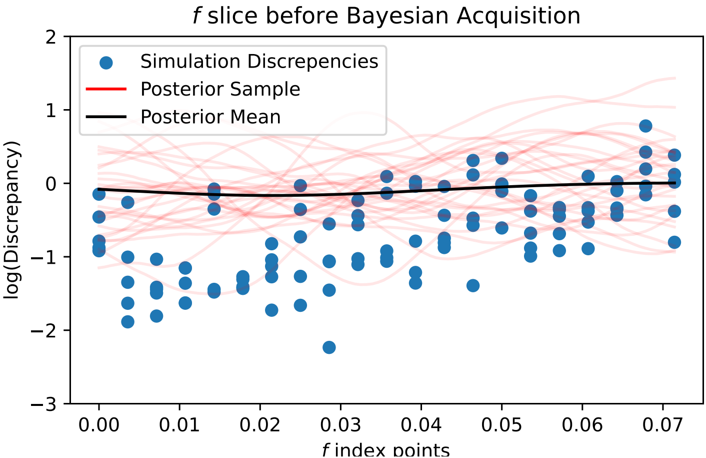


λ slice after 200 Bayesian acquisitions
Simulation Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
λ index points
2
1
0
−1
−2
−3
0.00
0.02
0.04
0.06
0.08
0.10

How did it go?

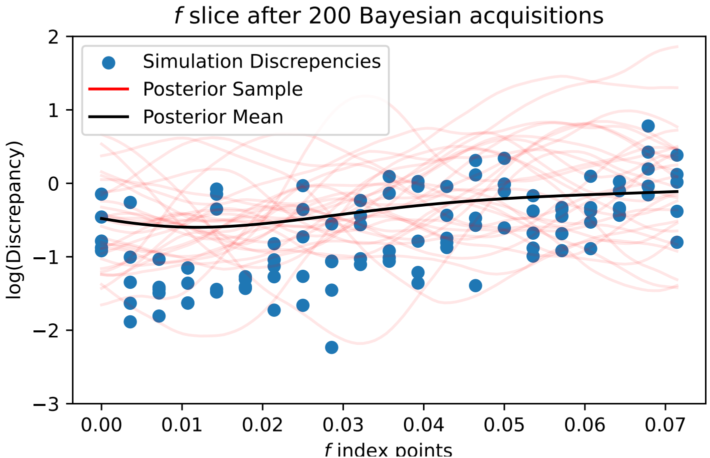

Final Sythetic Likelihood for λ Slice
1e−5
2.5
2.0
1.5
1.0
0.5
0.0
Synthetic likelihood
0.00
0.02
0.04
0.06
0.08
0.10
λ index points

How did it go?



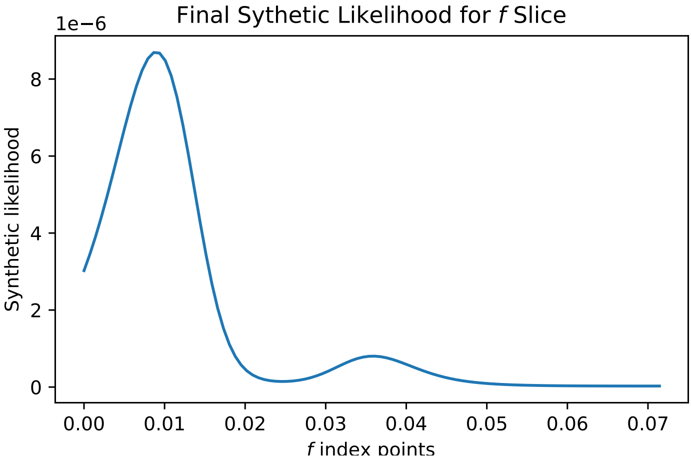
f slice before Bayesian Acquisition
Simulation Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
2
1
0
−1
−2
−3
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
f index points

How did it go?



f slice after 200 Bayesian acquisitions
Simulation Discrepencies
Posterior Sample
Posterior Mean
log(Discrepancy)
2
1
0
−1
−2
−3
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
f index points

How did it go?


Final Sythetic Likelihood for f Slice
1e−6
Synthetic likelihood
0
2
4
6
8
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
f index points

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 $\mathcal{D}(\boldsymbol{\theta})$
- Jumps where there is threshold/bifurcation behaviour
  - Student $t-$Process?

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Thanks to

- Eamon Conway
- Jennifer Flegg