Gaussian process accelerated ABC

Richard Wilkinson

University of Sheffield

Introduction

There are plenty of stochastic models which

- have unknown parameters
- are computationally expensive
- have unknown likelihood function
- are imperfect

Introduction

There are plenty of stochastic models which

- have unknown parameters
- are computationally expensive
- have unknown likelihood function
- are imperfect

E.g. Cellular Potts model for a human colon crypt

- agent-based models, with proliferation, differentiation and migration of cells
- stem cells generate a compartment of transient amplifying cells that produce colon cells.
- want to infer number of stem cells by comparing patterns with real data

Each simualtion takes $\sim$ 1 hour - efficient sampling will take us only so far...

Surrogate/Meta-modelling Emulation

Code uncertainty

For complex simulators, run times might be long, ruling out brute-force approaches such as Monte Carlo methods.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, f(\theta_i))\}_{i=1,\ldots,N}$$

- If θ is not in the ensemble, then we are uncertain about the value of $f(\theta)$.

Code uncertainty

For complex simulators, run times might be long, ruling out brute-force approaches such as Monte Carlo methods.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, f(\theta_i))\}_{i=1,\ldots,N}$$

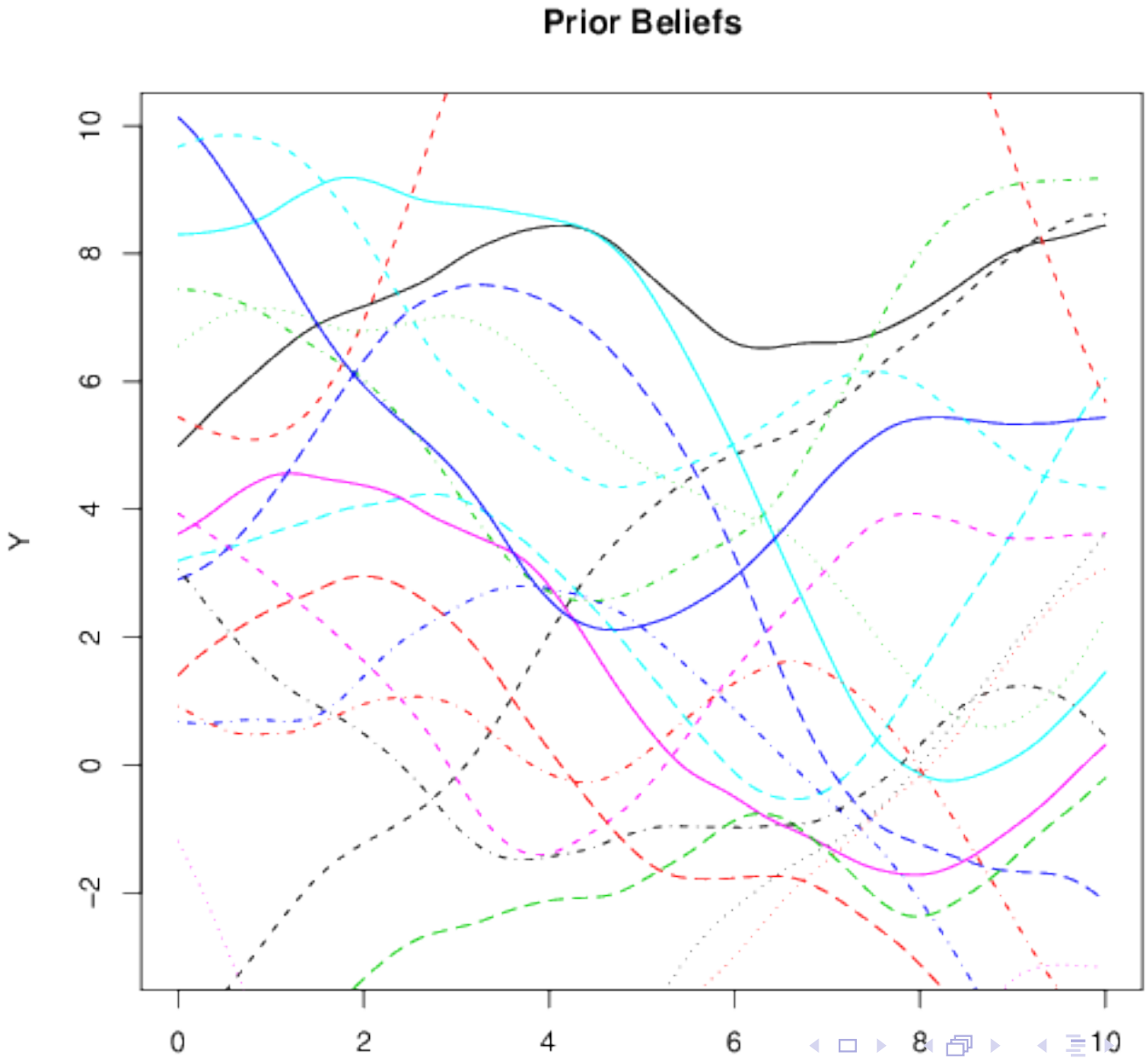
- If θ is not in the ensemble, then we are uncertain about the value of $f(\theta)$.

Idea: If the simulator is expensive, build a cheap model (*surrogate or emulator*) of it and use this in any analysis.

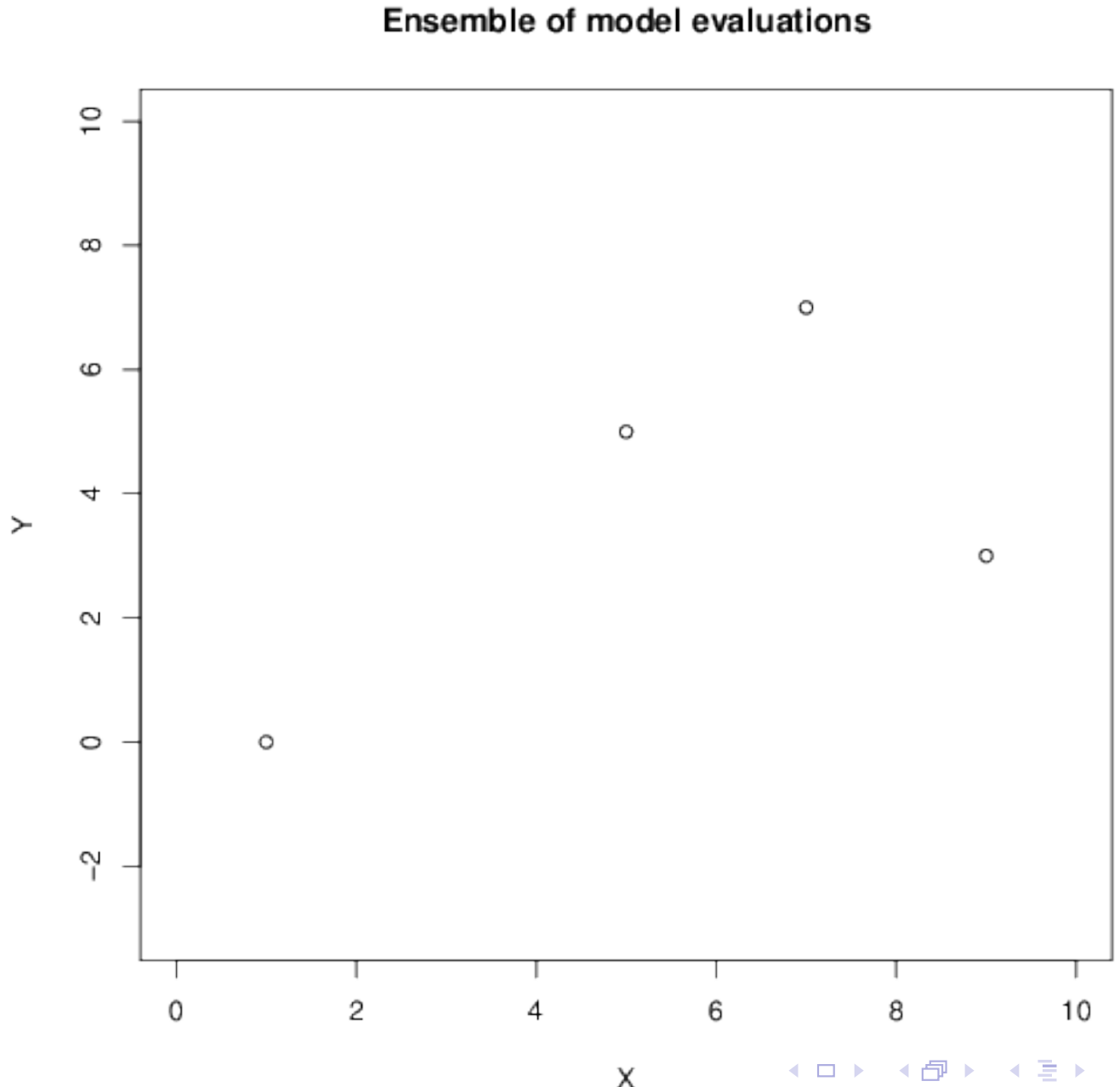
'a model of the model'

Gaussian Process Illustration

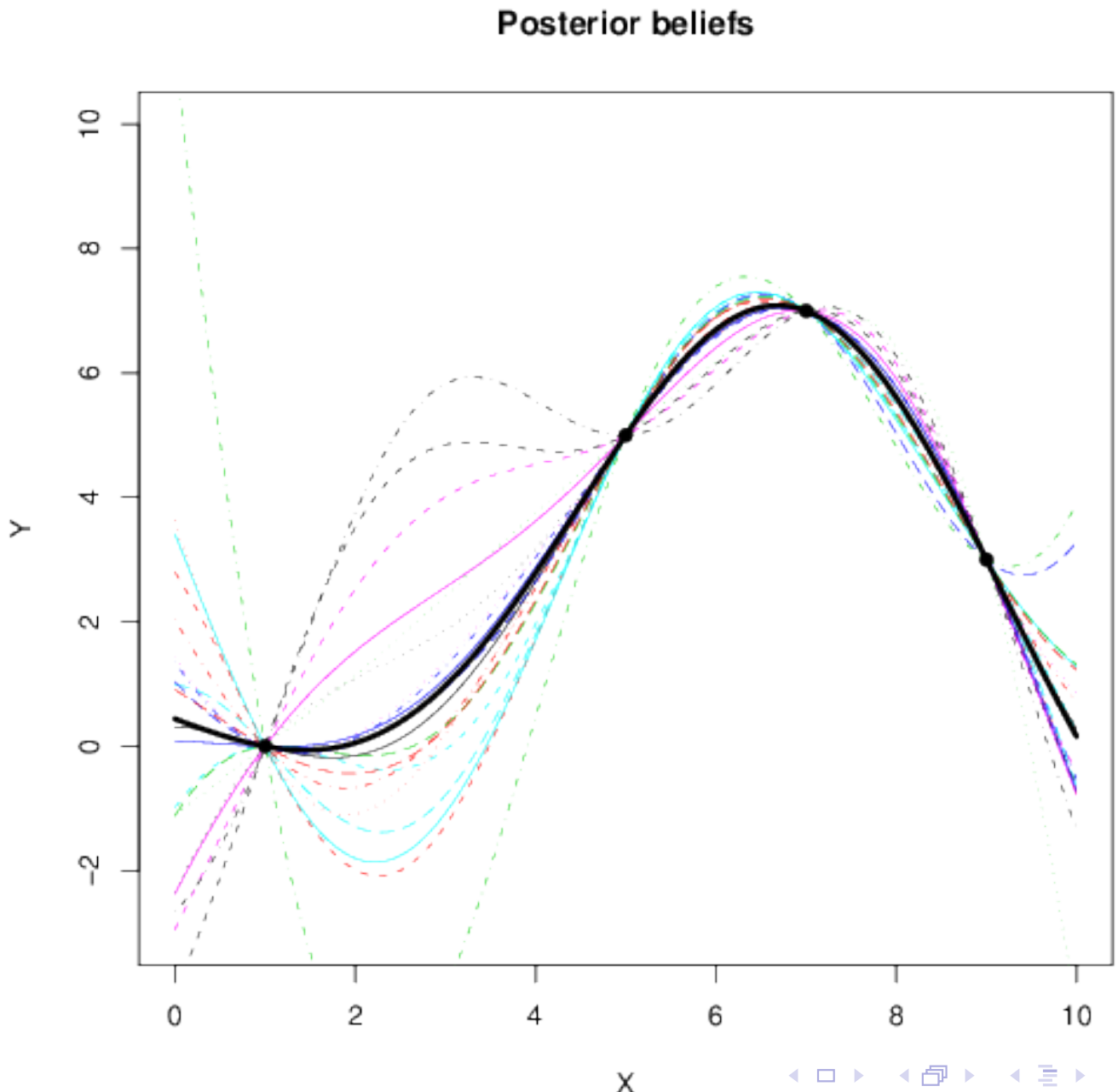
Zero mean



Gaussian Process Illustration



Gaussian Process Illustration



ABC

Inverse problems

- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- The inverse-problem: observe data D, estimate parameter values θ

Inverse problems

- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- The inverse-problem: observe data D, estimate parameter values θ

ABC algorithms are a collection of Monte Carlo methods used for calibrating simulators

- they do not require explicit knowledge of the likelihood function
- inference is done using simulation from the model (they are 'likelihood-free').

Rejection ABC

Uniform Rejection Algorithm

- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

Rejection ABC

Uniform Rejection Algorithm

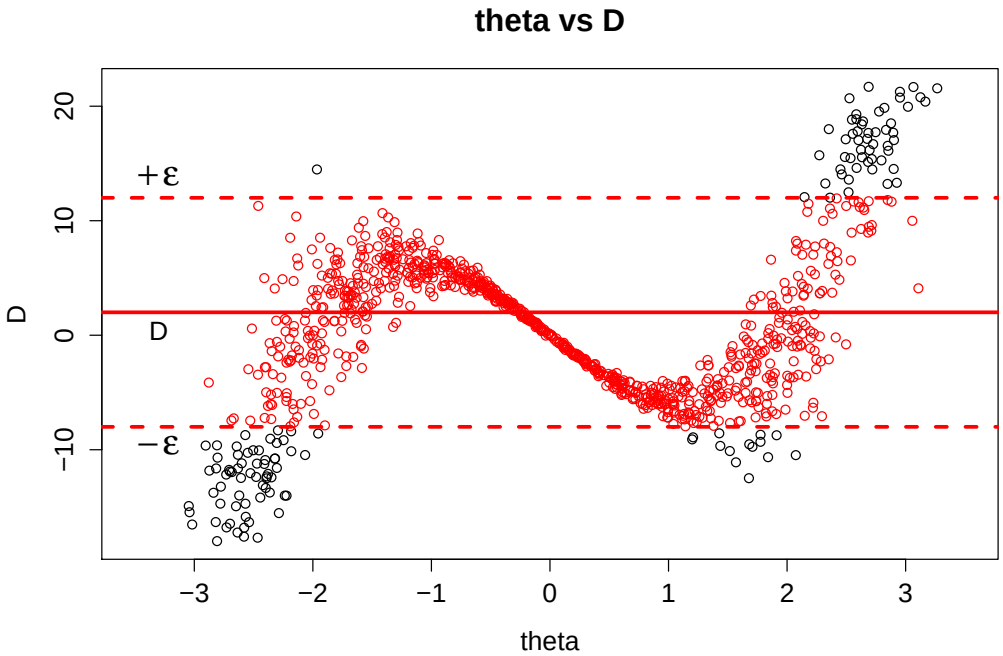
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

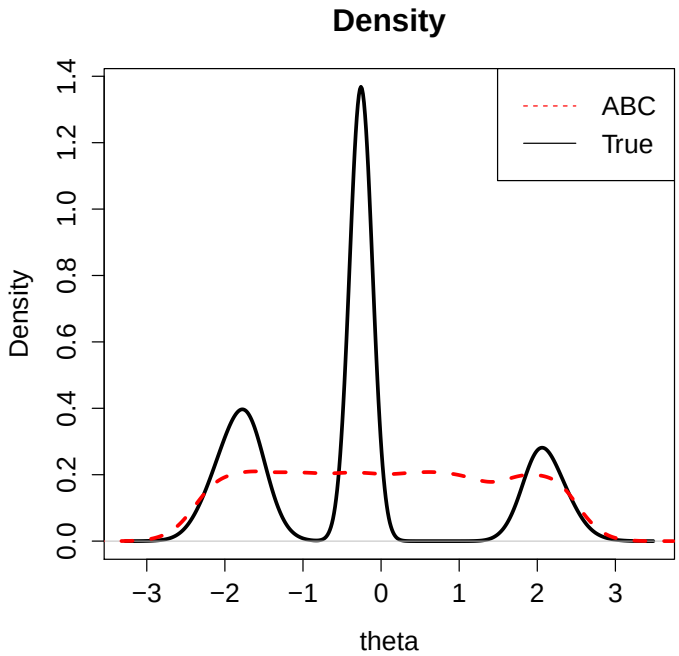
ϵ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

Rejection sampling is inefficient, but we can adapt other MC samplers such as MCMC and SMC.

$\epsilon = 10$

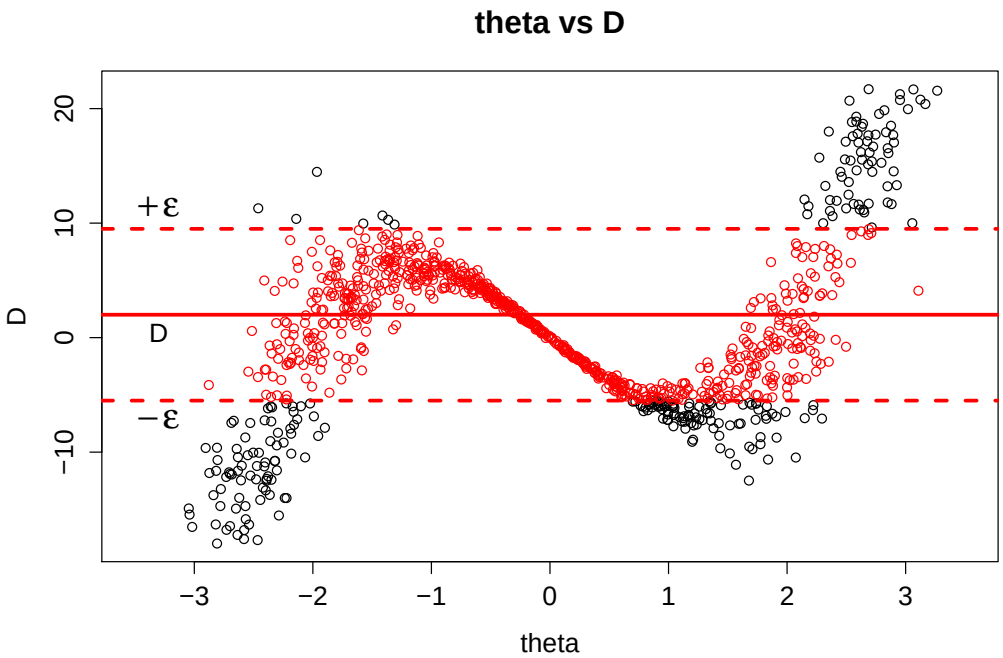


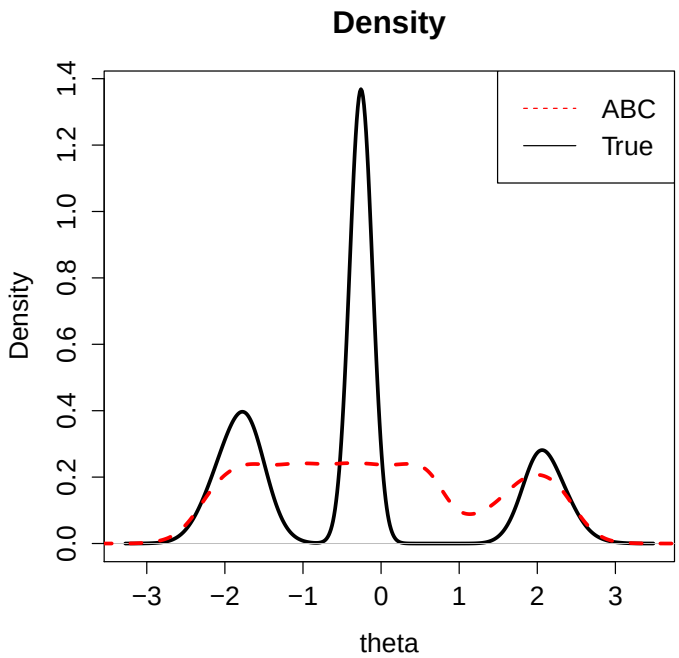


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

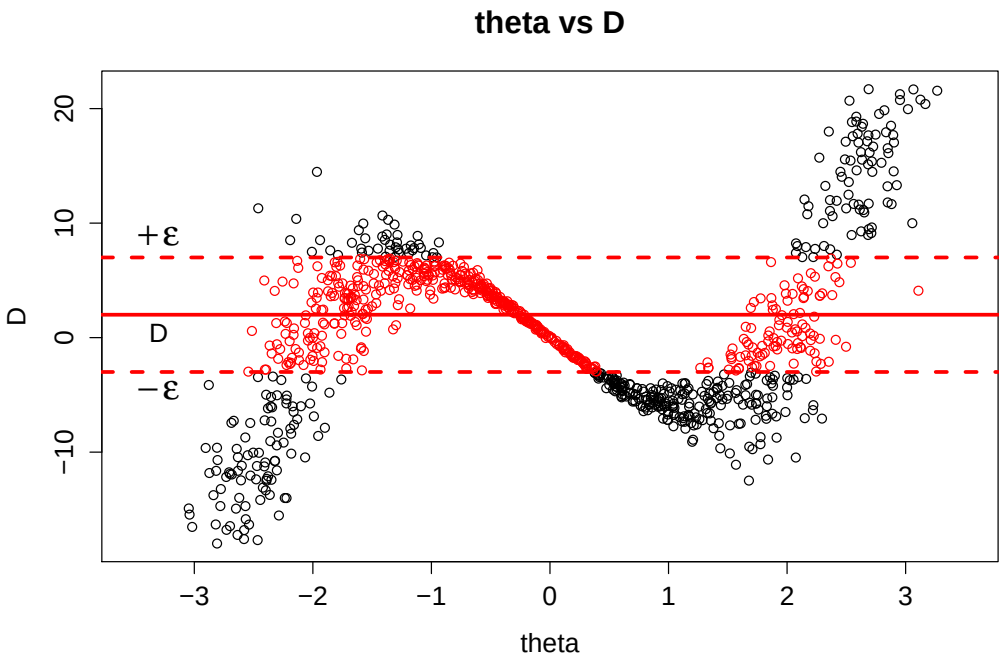
$$\rho(D, X) = |D - X|, \qquad D = 2$$

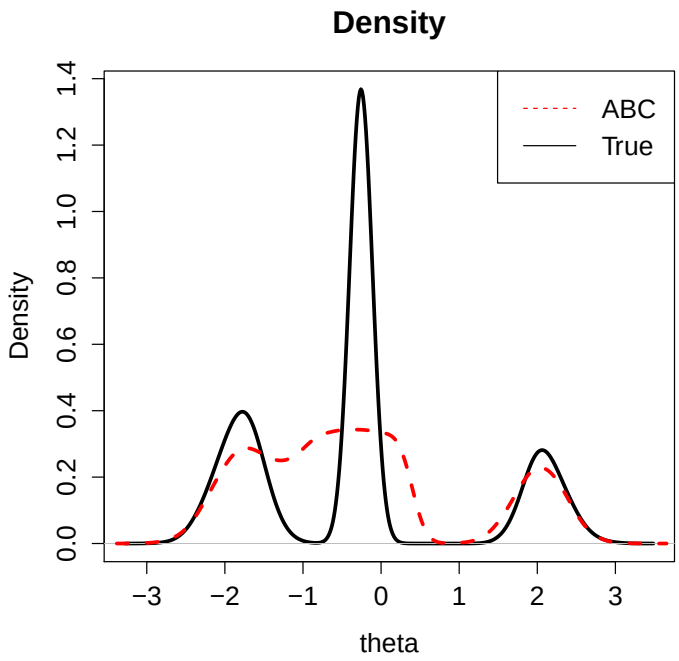
$\epsilon = 7.5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3

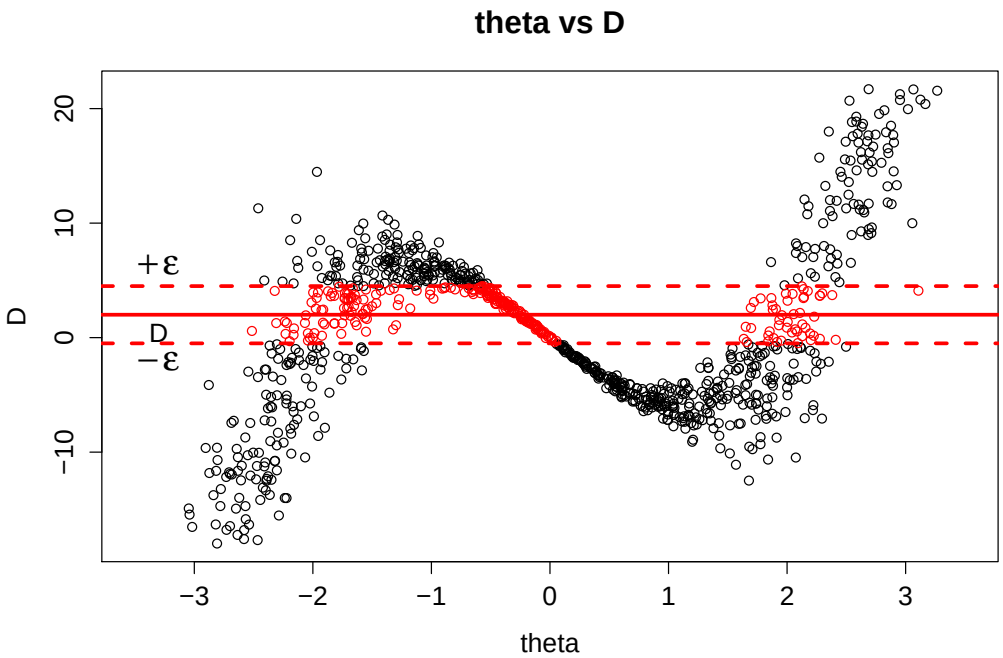

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

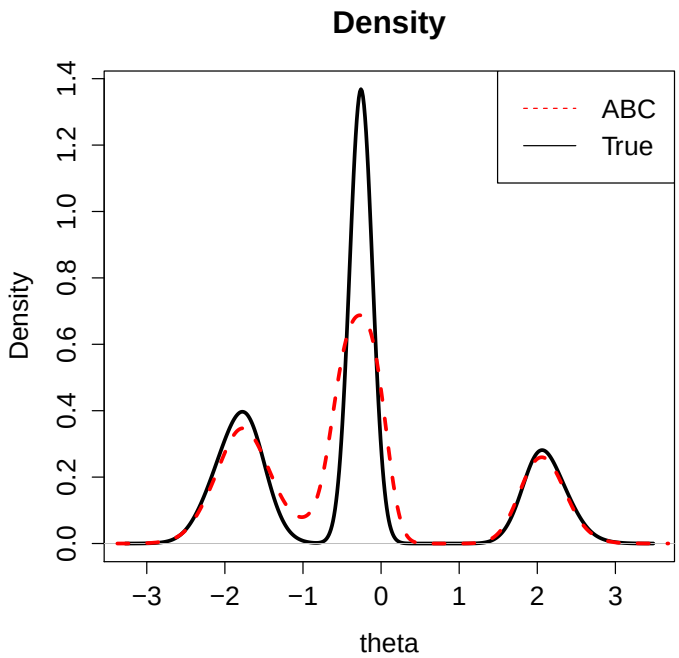
$\epsilon = 5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3

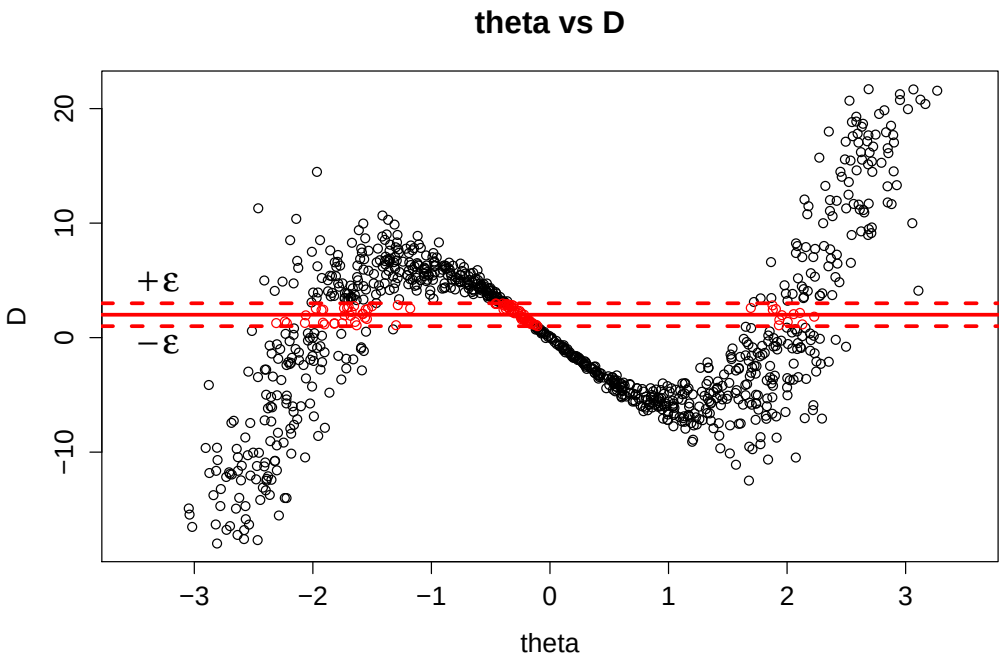

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

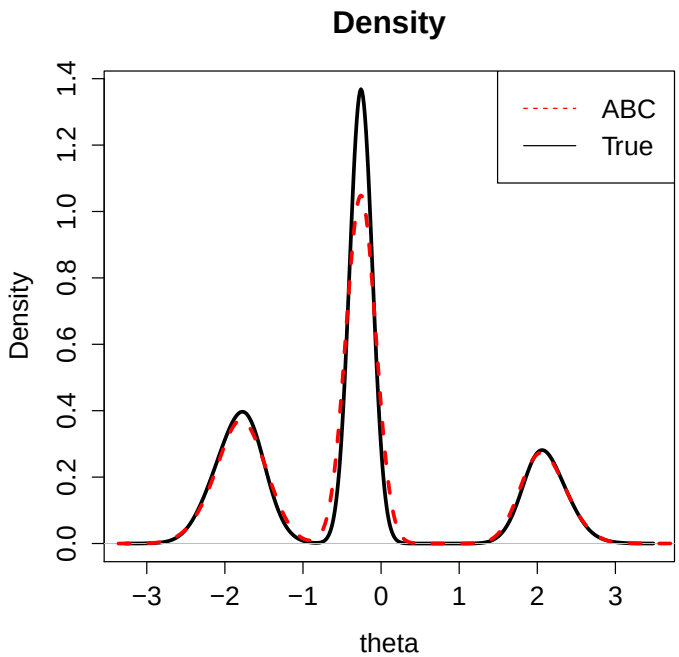
$\epsilon = 2.5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3


Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

$\epsilon = 1$


theta vs D
D
theta
+ε
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3


Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Surrogate ABC

Limitations of Monte Carlo methods

(Non approximate) Monte Carlo methods are generally guaranteed to succeed if we run them for long enough, but can require more simulation than is possible.

Limitations of Monte Carlo methods

(Non approximate) Monte Carlo methods are generally guaranteed to succeed if we run them for long enough, but can require more simulation than is possible.

Most MC methods

- sample naively - they don't learn from previous simulations.
- don't exploit known properties of the likelihood function, such as continuity
- sample randomly, rather than using careful design.

Limitations of Monte Carlo methods

(Non approximate) Monte Carlo methods are generally guaranteed to succeed if we run them for long enough, but can require more simulation than is possible.

Most MC methods

- sample naively - they don't learn from previous simulations.
- don't exploit known properties of the likelihood function, such as continuity
- sample randomly, rather than using careful design.

Idea: use surrogate modelling to speed up inference.

Surrogate ABC

- Wilkinson 2014
- Meeds and Welling 2014
- Gutmann and Corander 2015
- Strathmann, Sejdinovic, Livingstone, Szabo, Gretton 2015
- $\vdots$

With obvious influence from emulator community (e.g. Sacks, Welch, Mitchell, and Wynn 1989, Kennedy and O'Hagan 2001)

Constituent elements:

- Target of approximation
- Aim of inference and inference scheme
- Choice of surrogate/emulator
- Training/acquisition rule

$\exists$ a relationship to probabilistic numerics

Wood 2010

Wood 2010 introduced a synthetic likelihood

$$\pi(D|\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$

where μ_θ and Σ_θ are the mean and covariance of the simulator output when run at θ, and plugged this into an MCMC sampler.

Wood 2010

Wood 2010 introduced a synthetic likelihood

$$\pi(D|\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$

where μ_θ and Σ_θ are the mean and covariance of the simulator output when run at θ, and plugged this into an MCMC sampler.

- This suggested modelling dependence on θ to mitigate the cost

> *[...] the forward model may exhibit regularity in its dependence on the parameters of interest[...]. Replacing the forward model with an approximation or "surrogate"* **decouples** *the required number of forward model evaluations from the length of the MCMC chain, and thus can vastly reduce the overall cost of interence. Conrad et al. 2015*

Example: Ricker Model

Wood 2010

The Ricker model is one of the prototypic ecological models.

- used to model the fluctuation of the observed number of animals in some population over time
- It has complex dynamics and likelihood, despite its simple mathematical form.

Ricker Model

- Let N_t denote the number of animals at time t.

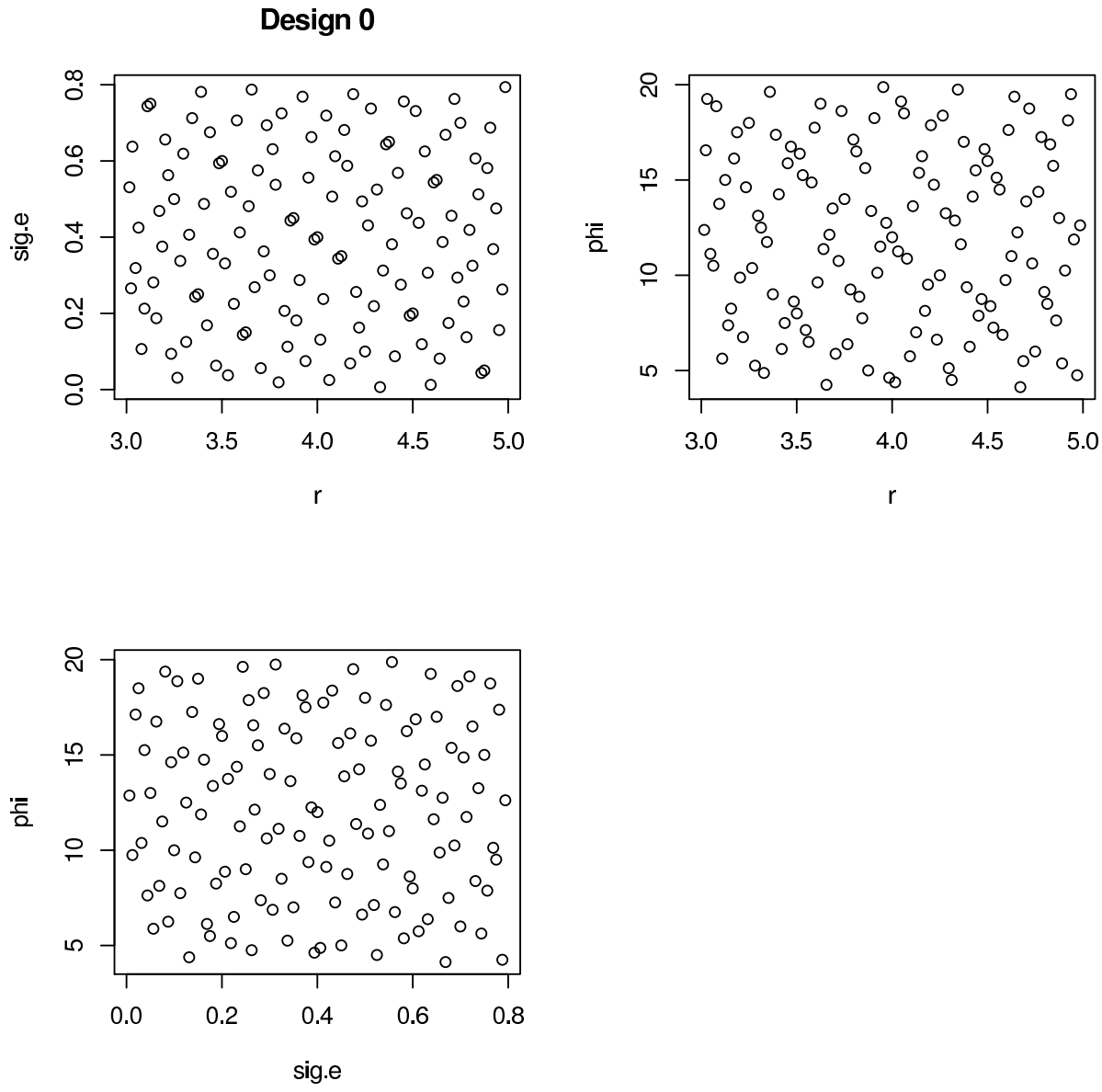
$$N_{t+1} = rN_t e^{-N_t + e_r}$$

where e_t are independent $N(0, \sigma_e^2)$ process noise

- Assume we observe counts y_t where

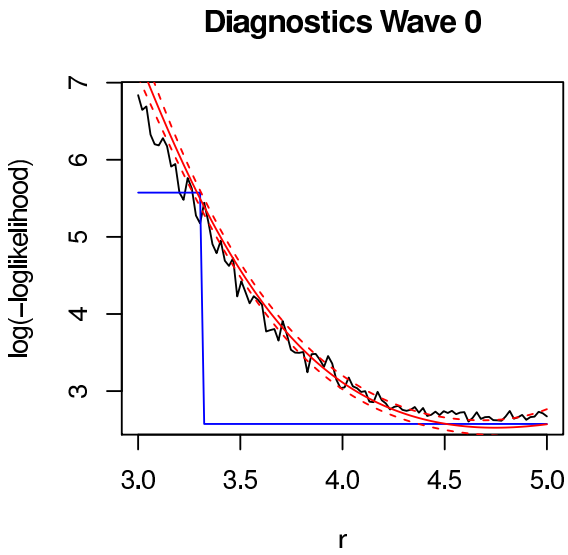
$$y_t \sim Po(\phi N_t)$$

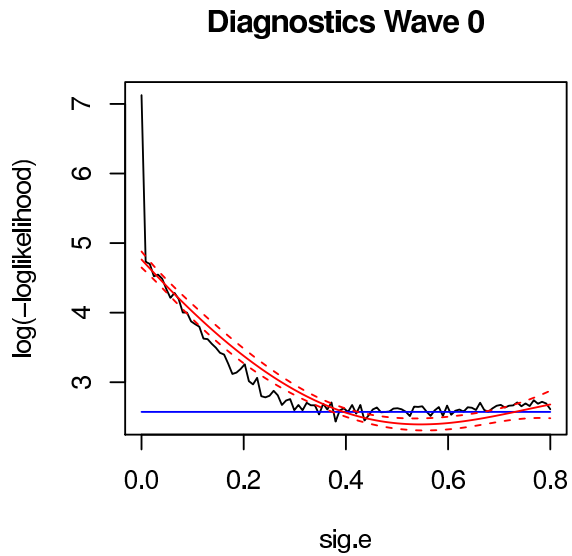
Results - Design 1 - 128 pts

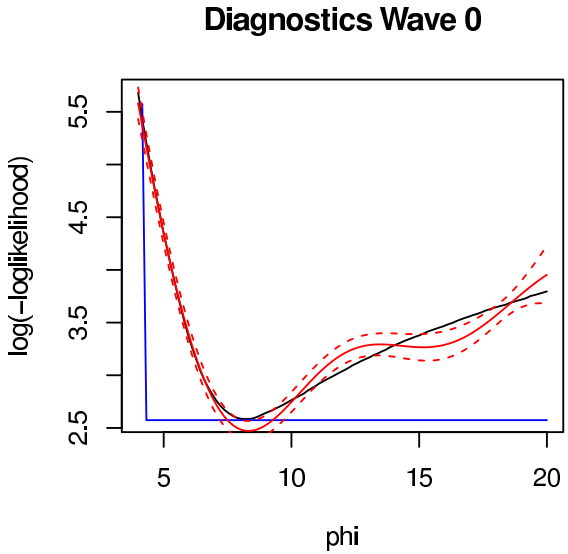


Diagnostics for GP 1 modelling $\log(-\log l(\theta))$

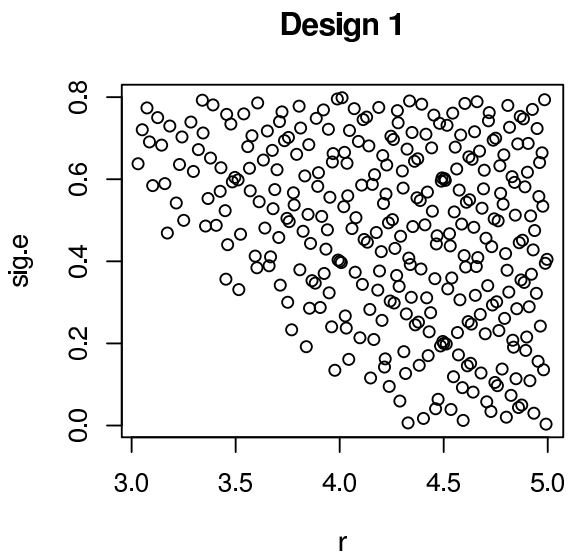
Threshold = 5.6

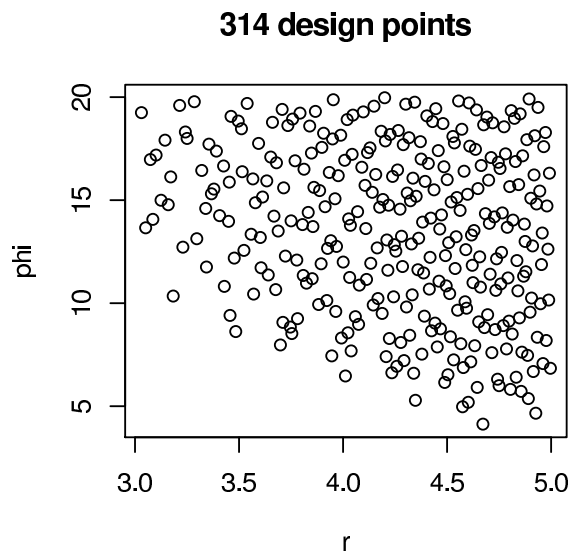


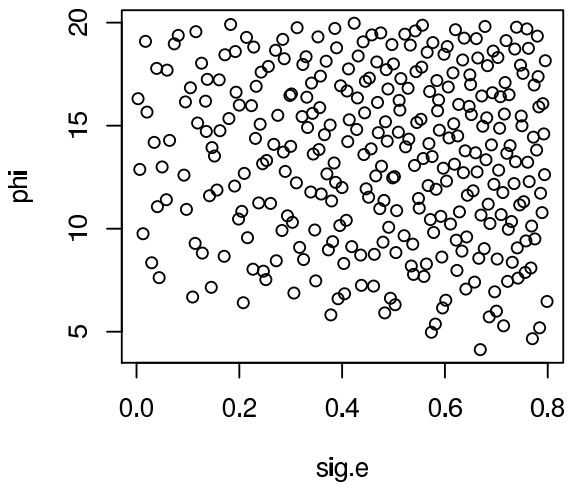




Results - Design 2 - 314 pts - 38% of space implausible

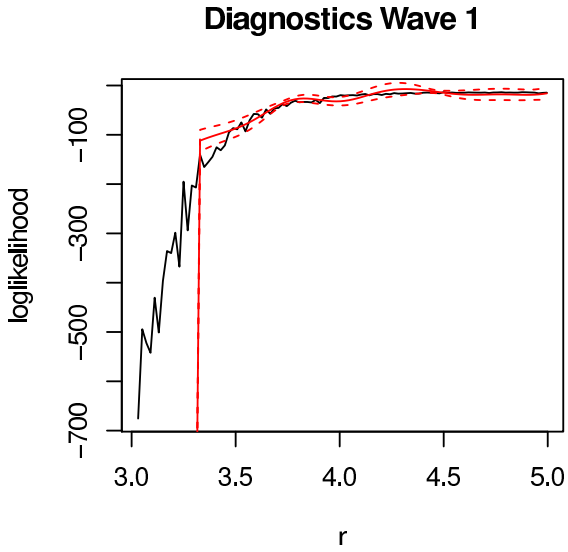


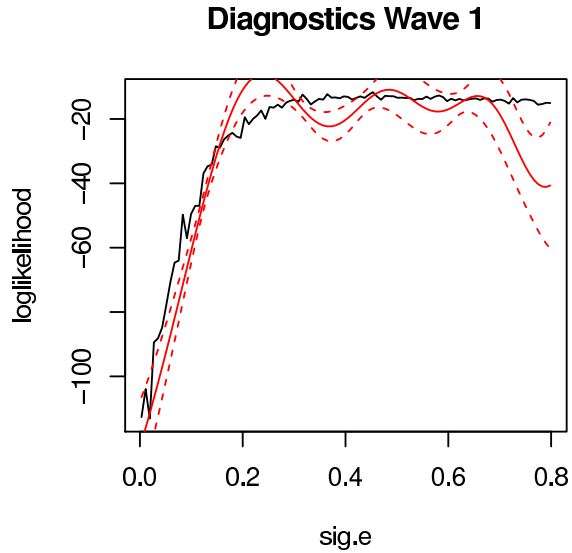


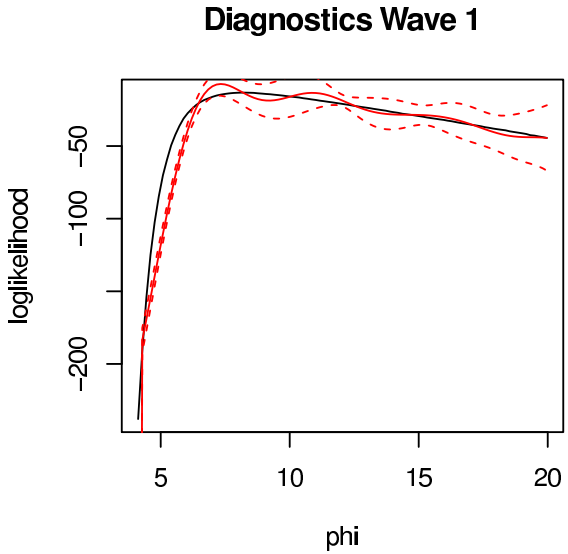


Diagnostics for GP 2 modelling log $l(\theta)$

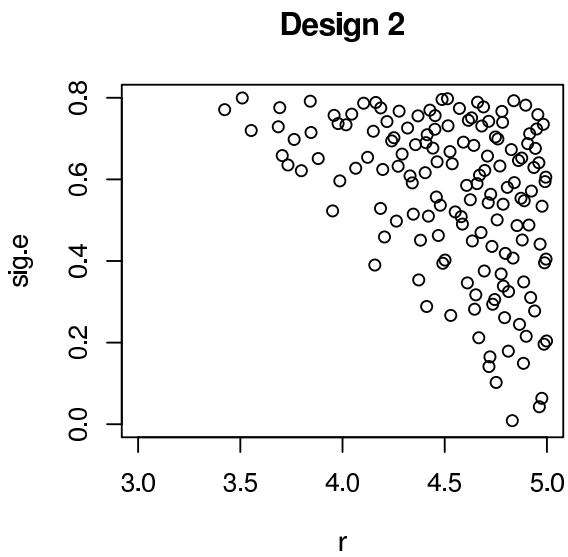
threshold = -21.8

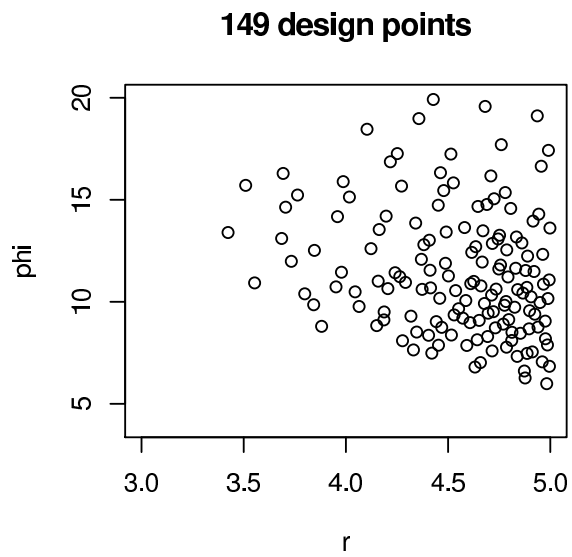


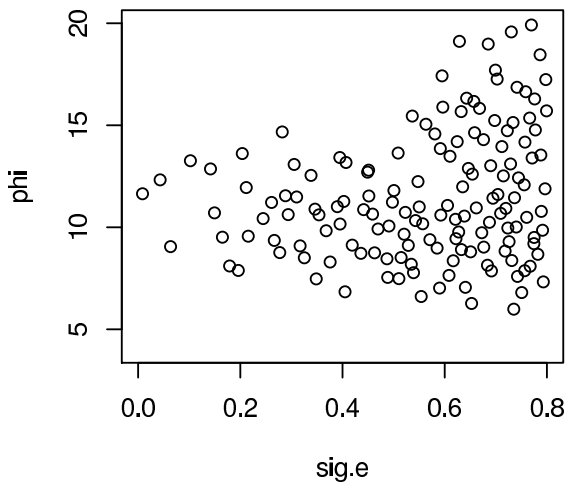




Design 3 - 149 pts - 62% of space implausible

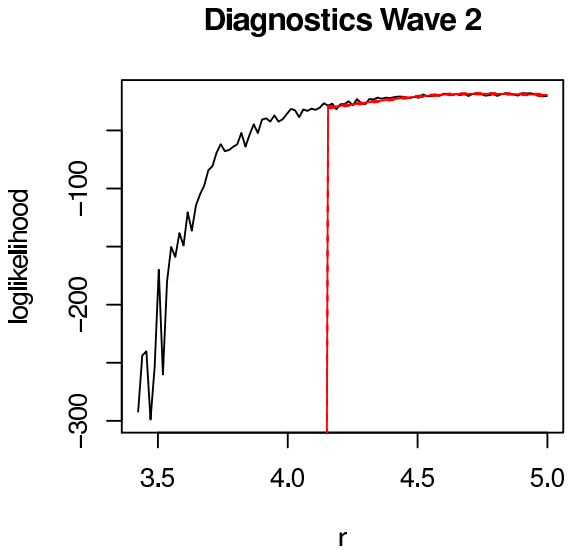


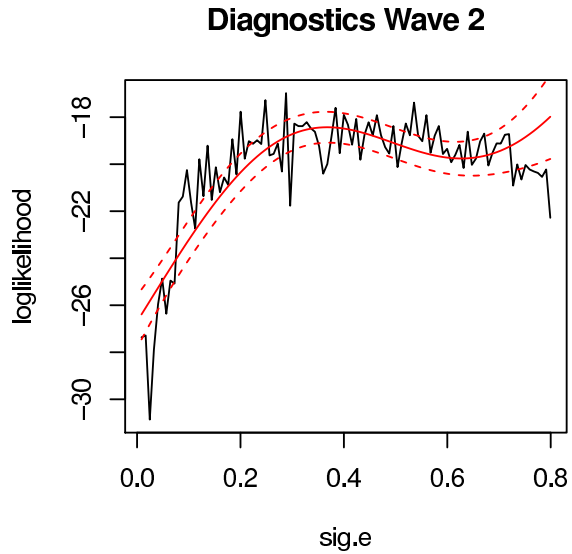


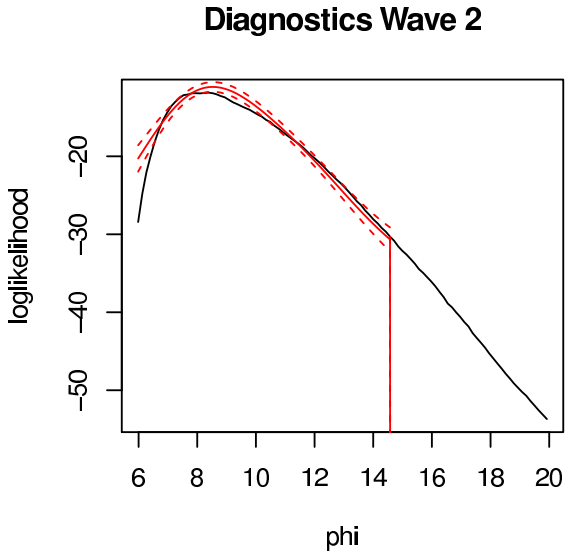


Diagnostics for GP 3 modelling $\log l(\theta)$

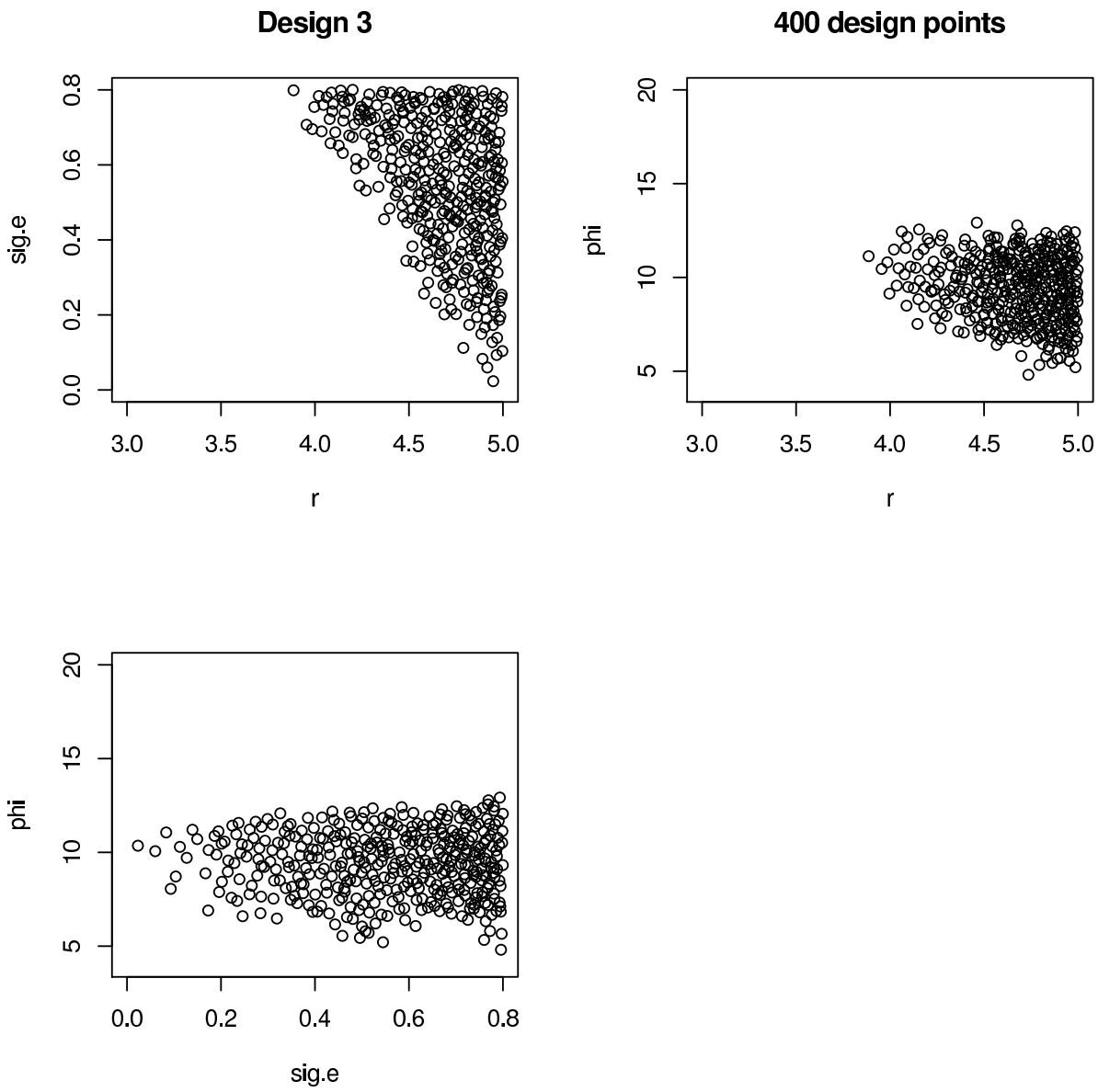
Threshold = -20.7





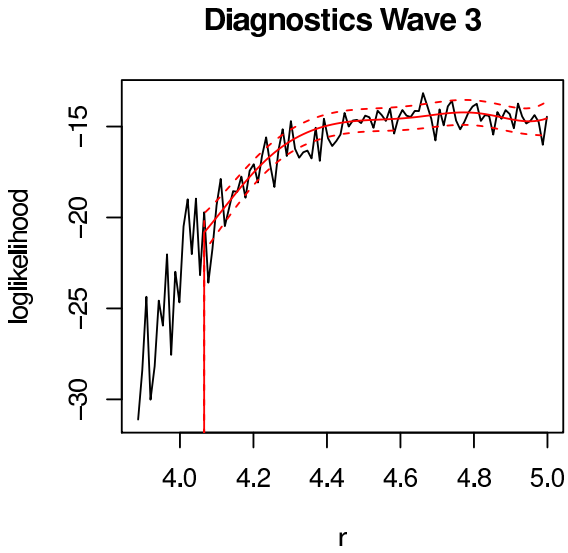


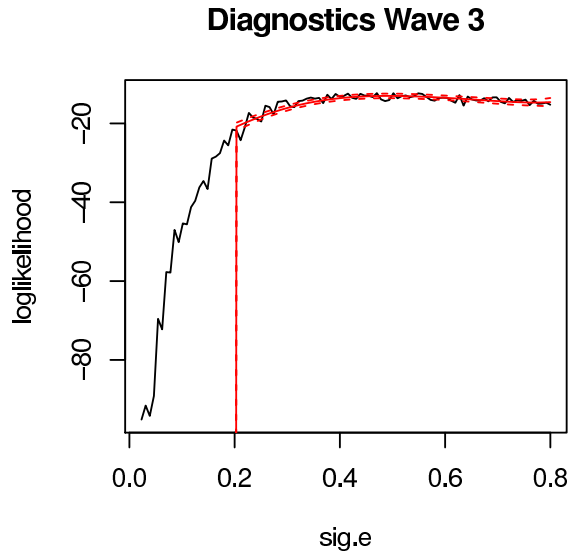
Design 4 - 400 pts - 95% of space implausible

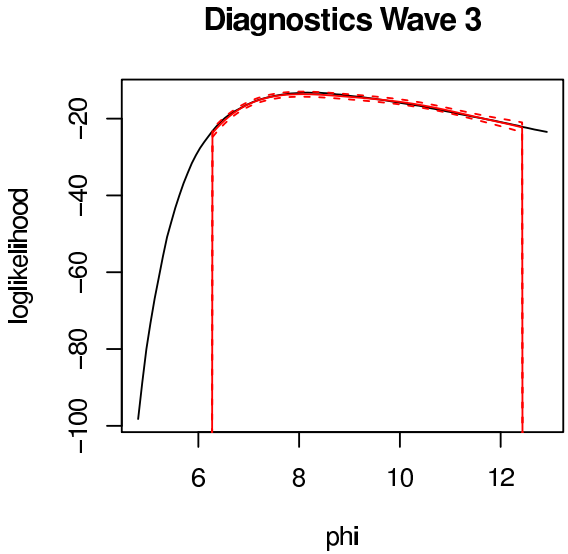


Diagnostics for GP 4 modelling log $l(\theta)$

Threshold = -16.4

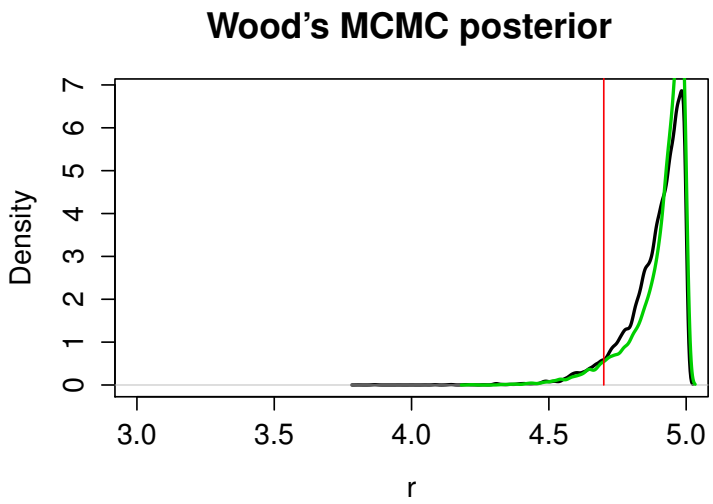


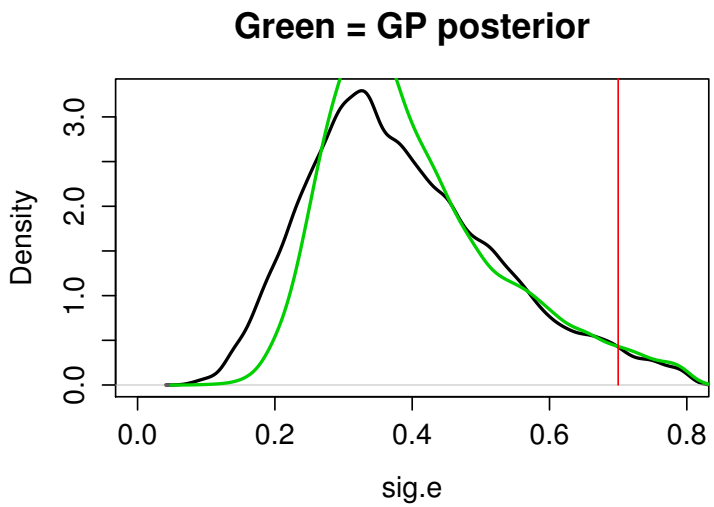


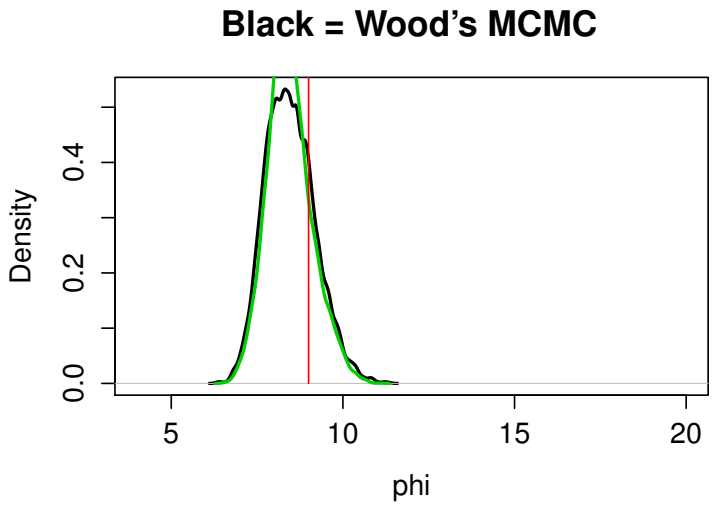


MCMC Results

Comparison with Wood 2010, synthetic likelihood approach







Computational details

- The Wood MCMC method used $10^5 \times 500$ simulator runs
- The GP code used $(128 + 314 + 149 + 400) = 991 \times 500$ simulator runs
  - 1/100th of the number used by Wood's method.

By the final iteration, the Gaussian processes had ruled out over 98% of the original input space as implausible,

- the MCMC sampler did not need to waste time exploring those regions.

Target of approximation for the surrogate

- Simulator output within synthetic likelihood (Meeds et al 2014) e.g.

$$\mu_\theta = \mathbb{E}f(\theta) \qquad \text{and} \qquad \Sigma_\theta = \mathbb{V}\text{ar}f(\theta)$$

- (ABC) Likelihood type function (Wilkinson 2014)

$$\begin{aligned} L_{ABC}(\theta) &= \mathbb{E}_{X|\theta} K_\epsilon[\rho(T(D), T(X))] \\ &\equiv \mathbb{E}_{X|\theta} \pi_\epsilon(D|X) \approx \frac{1}{N}\sum_{i=1}^{N} \pi(D|X_i) \end{aligned}$$

- Discrepancy function (Gutmann and Corander, 2015), for example

$$J(\theta) = \mathbb{E}\rho(S(D), S(X))$$

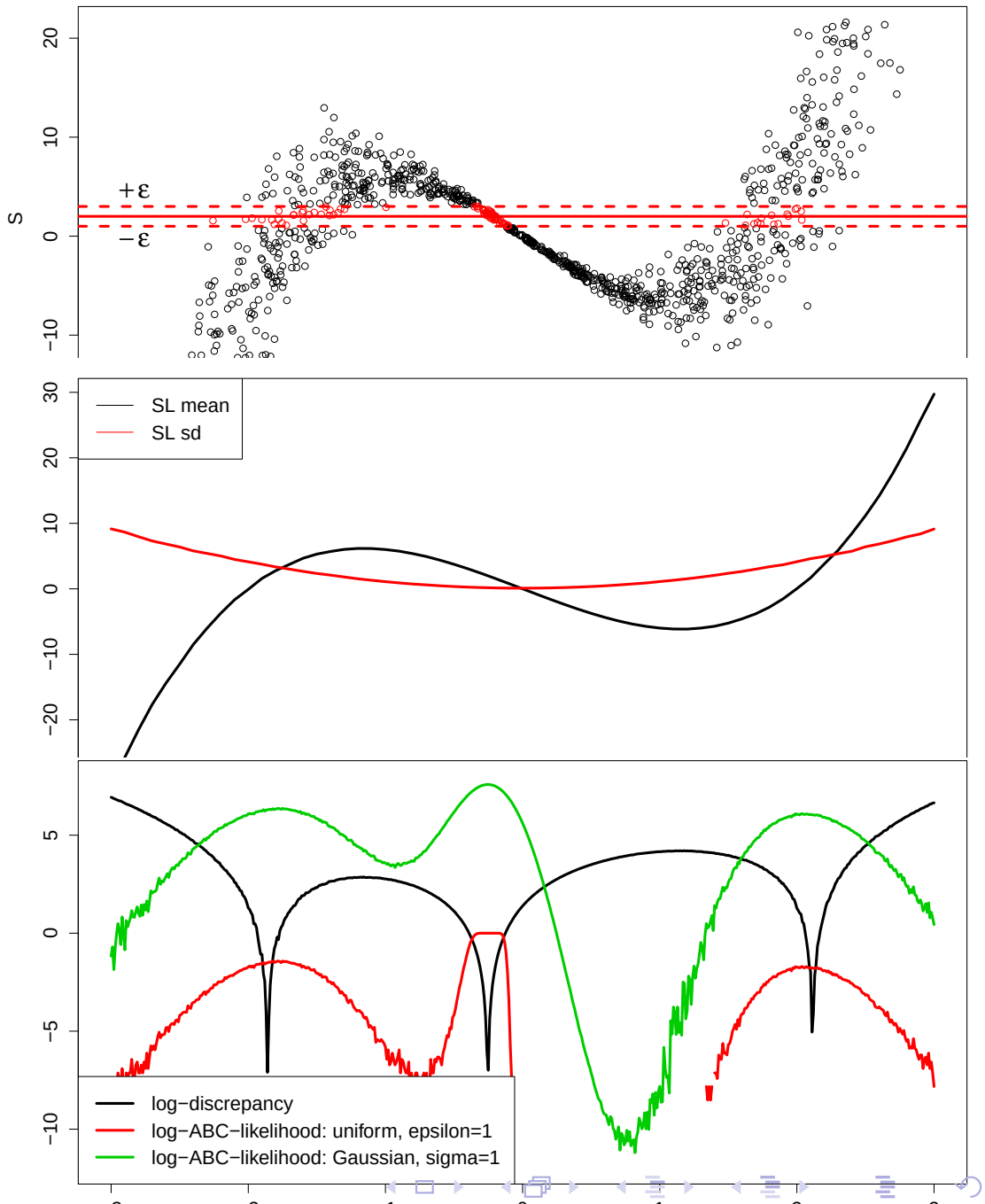
- Gradients (Strathmann et al 2015)

The difficulty of each approach depends on smoothness, dimension, focus etc.

$$S \sim N(2(\theta+2)\theta(\theta-2), 0.1+\theta^2)$$

Synthetic likelihood:

ABC likelihood and discrepancy:



Choice of surrogate: sequential history matching approach

Wilkinson 2014

The log-likelihood $\ell(\theta) = \log L(\theta)$ typical ranges across too wide a range of values, consequently, most models struggle to accurately approximate the log-likelihood across the entire parameter space.

Choice of surrogate: sequential history matching approach

Wilkinson 2014

The log-likelihood $\ell(\theta) = \log L(\theta)$ typical ranges across too wide a range of values, consequently, most models struggle to accurately approximate the log-likelihood across the entire parameter space.

- Introduce waves of history matching.
- In each wave, build a GP model that can rule out regions of space as implausible.
  - Say θ implausible if $\ell(\theta) < \ell(\hat{\theta}) - T$ for some conservative threshold T.
  - Ruling θ to be implausible is to set $\pi(\theta|y) = 0$

Choice of surrogate: sequential history matching approach

Wilkinson 2014

The log-likelihood $\ell(\theta) = \log L(\theta)$ typical ranges across too wide a range of values, consequently, most models struggle to accurately approximate the log-likelihood across the entire parameter space.

- Introduce waves of history matching.
- In each wave, build a GP model that can rule out regions of space as implausible.
  - Say θ implausible if $\ell(\theta) < \ell(\hat{\theta}) - T$ for some conservative threshold T.
  - Ruling θ to be implausible is to set $\pi(\theta|y) = 0$

We are uncertain about $\ell(\cdot)$ and $\hat{\theta}$, so decide that θ is implausible if

$$\mathbb{P}(\tilde{\ell}(\theta) < \max_{\theta_i} \ell(\theta_i) - T) \leq 0.001$$

where $\tilde{\ell}(\theta)$ is the GP model of $\log \pi(D|\theta)$

The choice of T is problem specific, and we can begin with a large T to ensure a conservative criterion.

Inference - controlling the error

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.

- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage

Inference - controlling the error

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
  - Hard to say anything about the error in the posterior
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage
  - under-utilizes the surrogate, sacrificing speed-up.

Inference - controlling the error

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
  - Hard to say anything about the error in the posterior
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage

Instead, Conrad *et al.* 2015 developed an intermediate approach that asymptotically samples from the exact posterior.

- proposes new θ - if uncertainty in surrogate prediction is such that it is unclear whether to accept or reject, then rerun simulator, else trust surrogate.

Inference - controlling the error

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
  - Hard to say anything about the error in the posterior
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage

Instead, Conrad *et al.* 2015 developed an intermediate approach that asymptotically samples from the exact posterior.

- proposes new θ - if uncertainty in surrogate prediction is such that it is unclear whether to accept or reject, then rerun simulator, else trust surrogate.

> *It is inappropriate to be concerned about mice when there are tigers abroad (Box 1976)*

Model discrepancy, ABC approximations, sampling errors etc may mean it is not worth worrying...

Design for calibration using Bayesian optimization

with James Hensman

Implausibility

When using emulators for history-matching and ABC, the aim is to accurately classify space as plausible or implausible by estimating the probability

$$p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta)$$

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$, based upon a GP model of the simulator or likelihood

$$f(\theta) \sim GP(m(\cdot), c(\cdot, \cdot))$$

Implausibility

When using emulators for history-matching and ABC, the aim is to accurately classify space as plausible or implausible by estimating the probability

$$p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta)$$

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$, based upon a GP model of the simulator or likelihood

$$f(\theta) \sim GP(m(\cdot), c(\cdot, \cdot))$$

The key determinant of emulator accuracy is the design used to train the GP

$$D_n = \{\theta_i, f(\theta_i)\}_{i=1}^N$$

Usual design choices are space filling designs

- Maximin latin hypercubes, Sobol sequences

Implausibility

When using emulators for history-matching and ABC, the aim is to accurately classify space as plausible or implausible by estimating the probability

$$p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta)$$

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$, based upon a GP model of the simulator or likelihood

$$f(\theta) \sim GP(m(\cdot), c(\cdot, \cdot))$$

The key determinant of emulator accuracy is the design used to train the GP

$$D_n = \{\theta_i, f(\theta_i)\}_{i=1}^N$$

Usual design choices are space filling designs

- Maximin latin hypercubes, Sobol sequences

Calibration doesn't need a global approximation to the simulator - this is wasteful

Entropic designs

Instead build a sequential design $\theta_1, \theta_2, \ldots$ using the current classification

$$p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta | D_n)$$

to guide the choice of design points

Entropic designs

Instead build a sequential design $\theta_1, \theta_2, \ldots$ using the current classification

$$p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta | D_n)$$

to guide the choice of design points
First idea: add design points where we are most uncertain

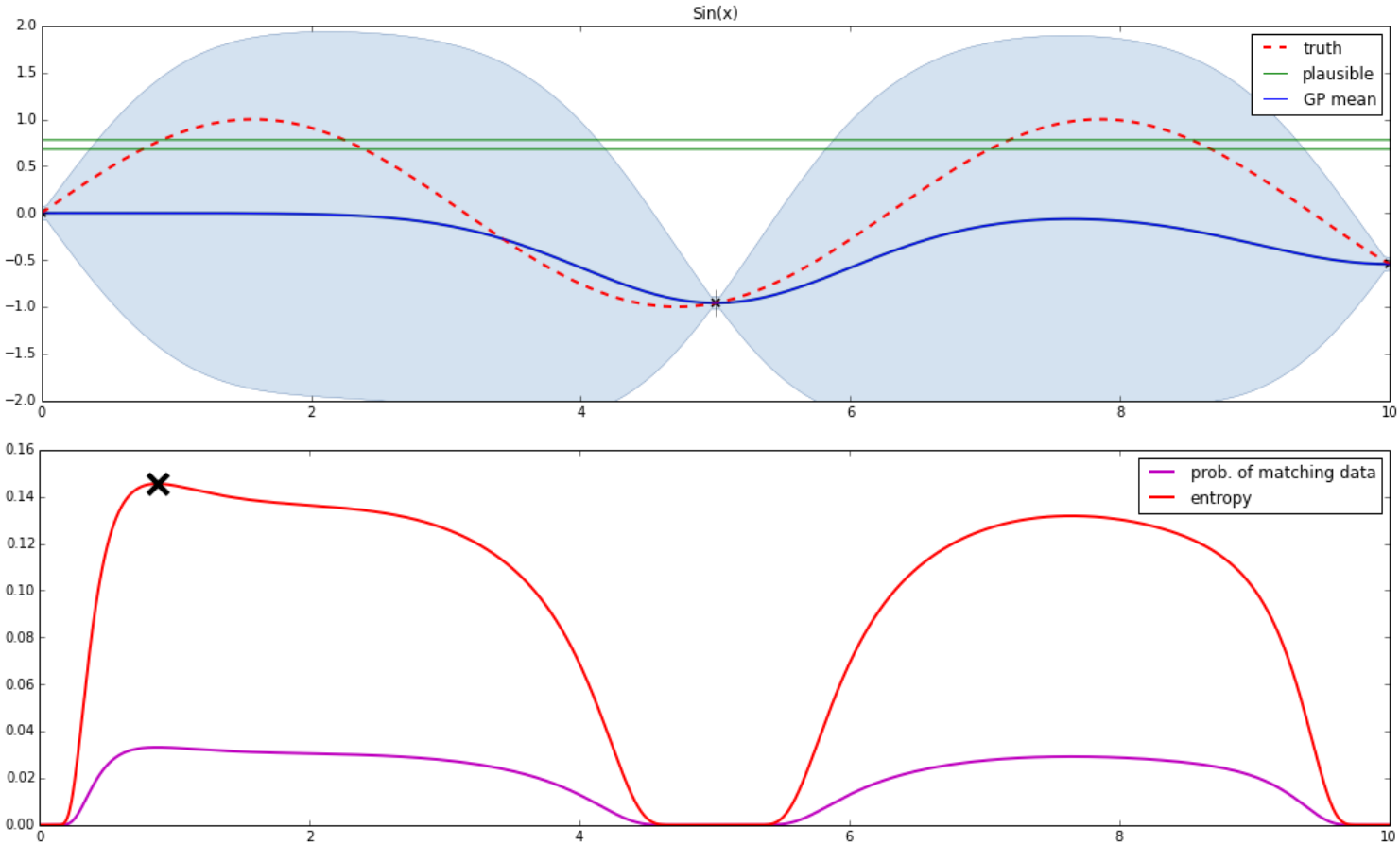
- The entropy of the classification surface is

$$E(\theta) = -p(\theta) \log p(\theta) - (1 - p(\theta)) \log(1 - p(\theta))$$

- Choose the next design point where we are most uncertain.

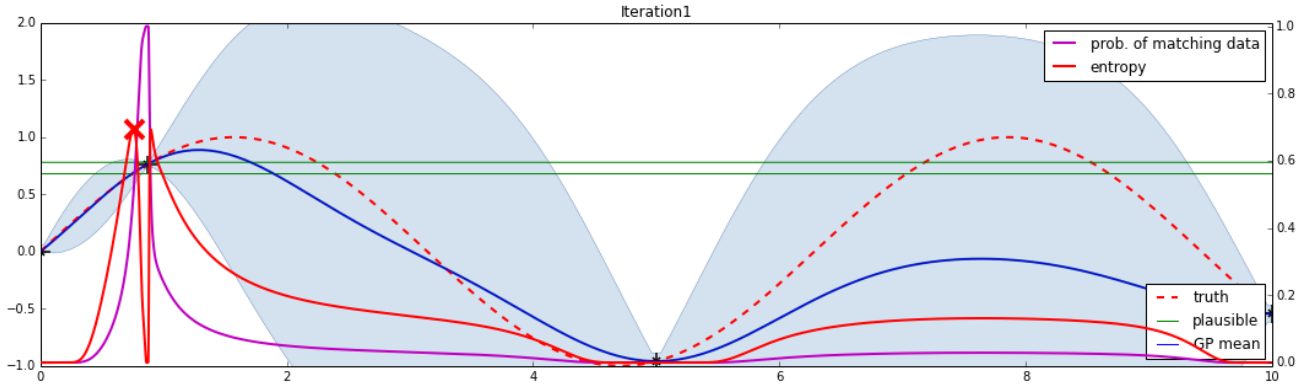
$$\theta_{n+1} = \arg\max E(\theta)$$

Toy 1d example $f(\theta) = \sin\theta$

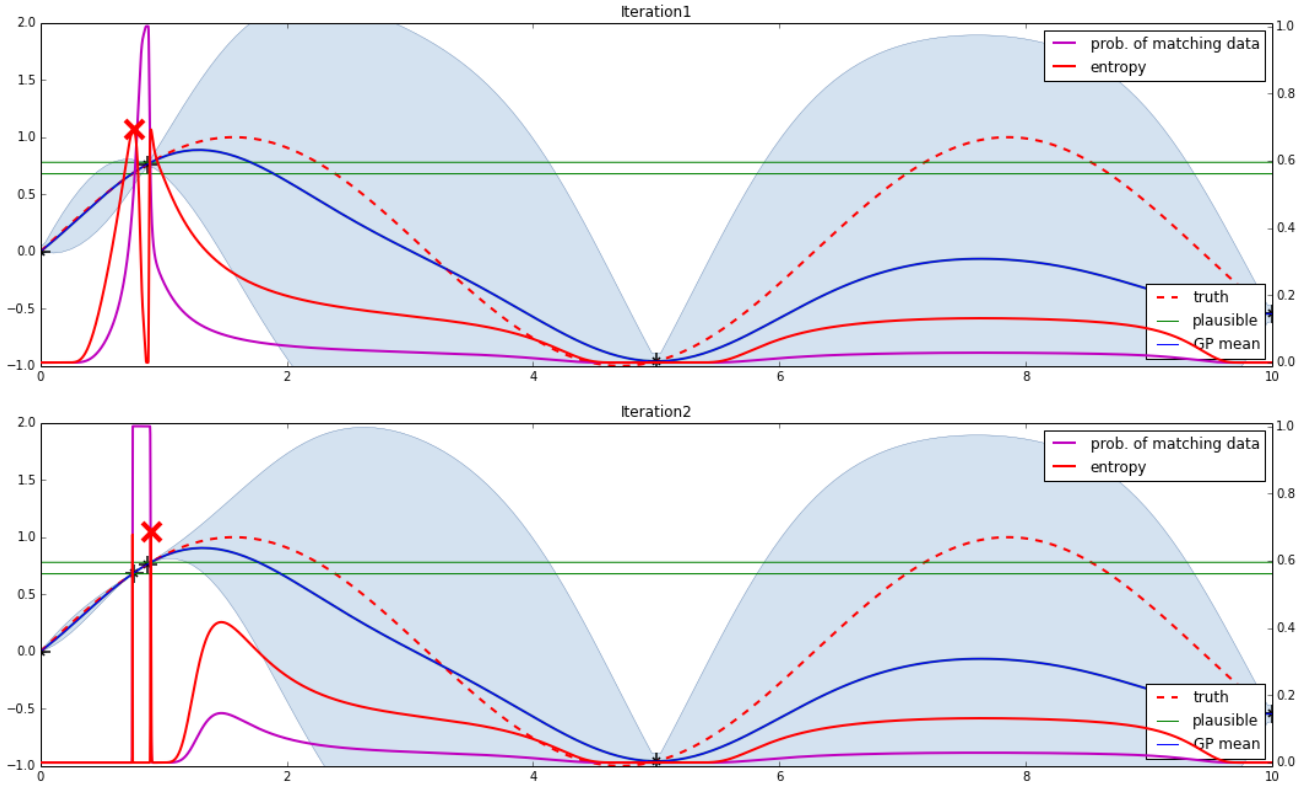


Add a new design point (simulator evaluation) at the point of greatest entropy

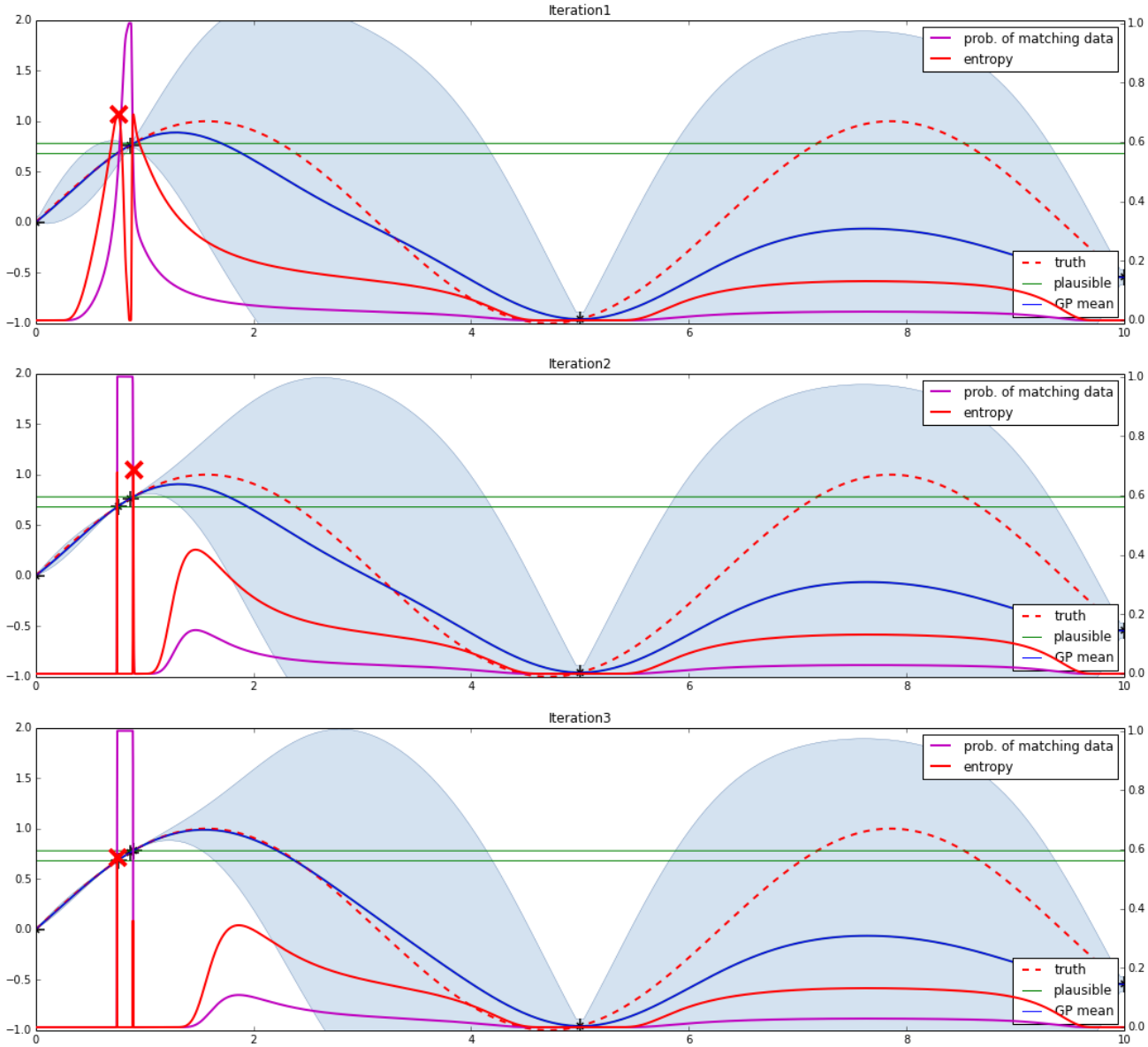
Toy 1d example $f(\theta) = \sin\theta$



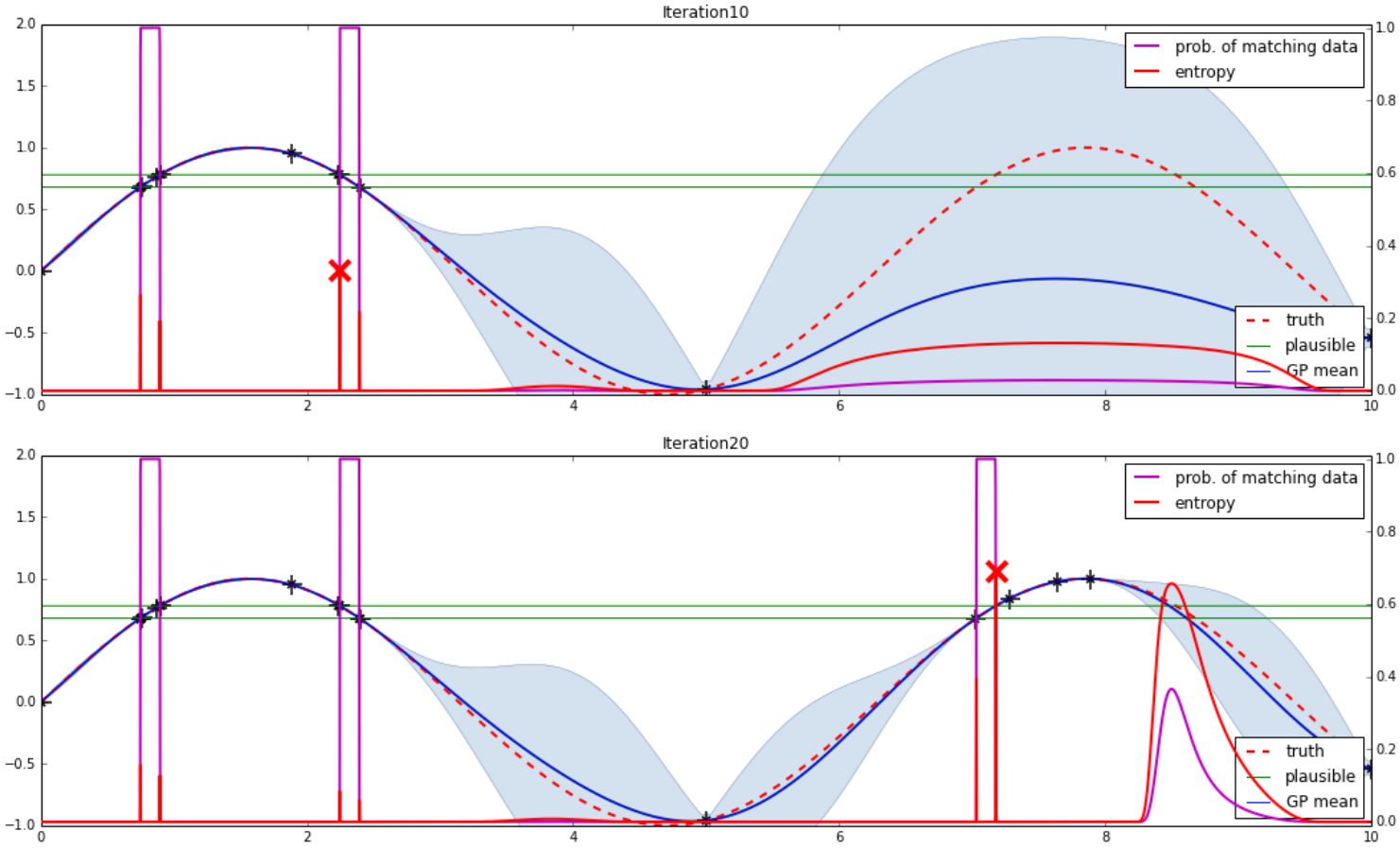
Toy 1d example $f(\theta) = \sin\theta$



Toy 1d example $f(\theta) = \sin\theta$



Toy 1d example $f(\theta) = \sin\theta$ - After 10 and 20 iterations



This criterion spends too long resolving points at the edge of the classification region.

- not enough exploration

Expected average entropy

Chevalier *et al.* 2014

Instead, we can find the average entropy of the classification surface

$$E_n = \int E(\theta) \mathrm{d}\theta$$

where n denotes it is based on the current design of size n.

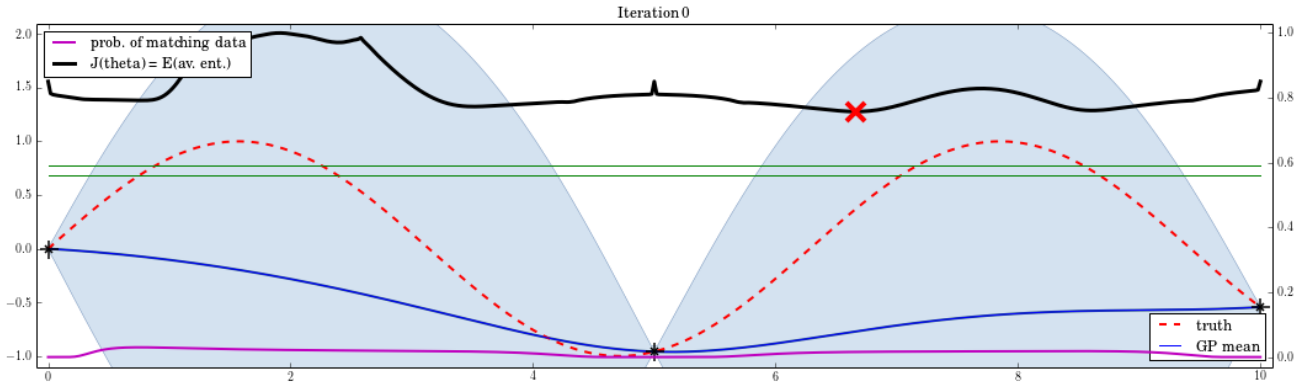
- Choose the next design point, θ_{n+1}, to minimise the expected average entropy

  $$\theta_{n+1} = \arg\min J_n(\theta)$$

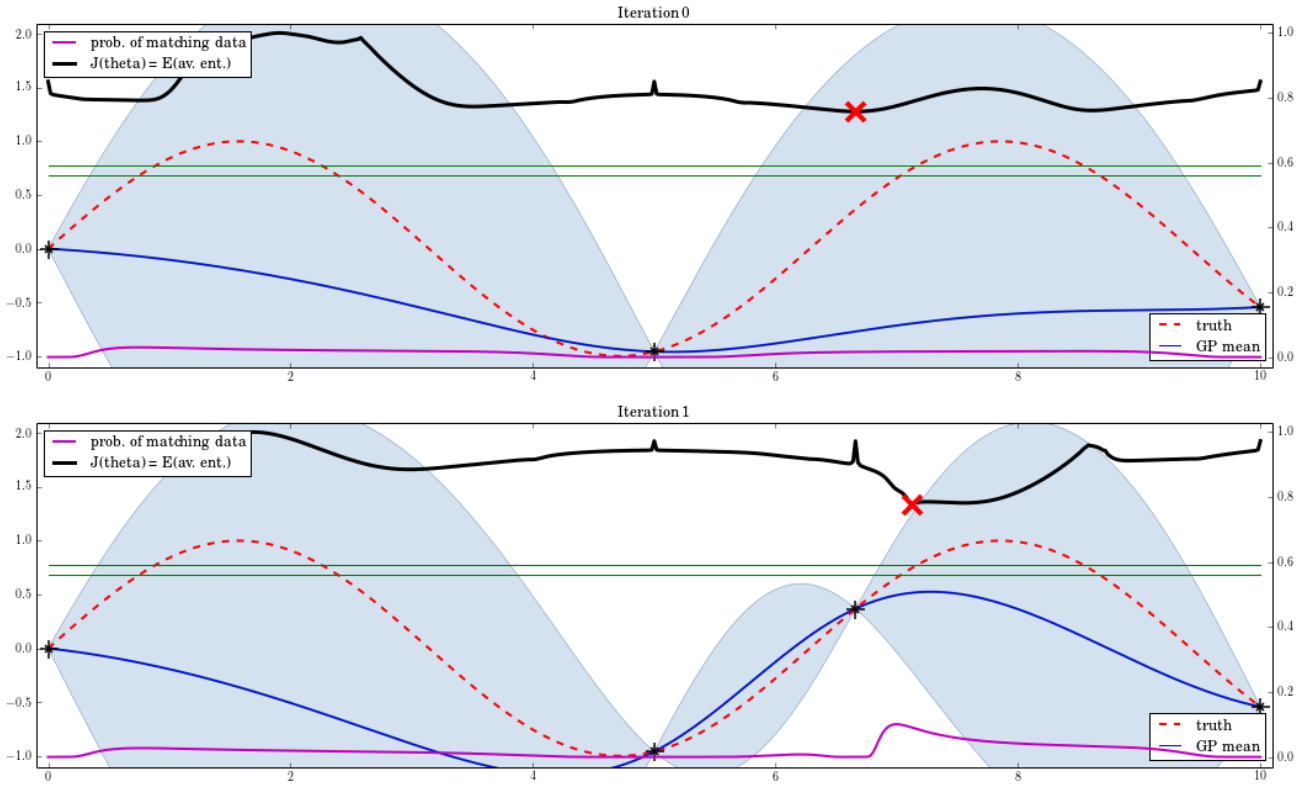
  where

  $$J_n(\theta) = \mathbb{E}(E_{n+1} | \theta_{n+1} = \theta)$$

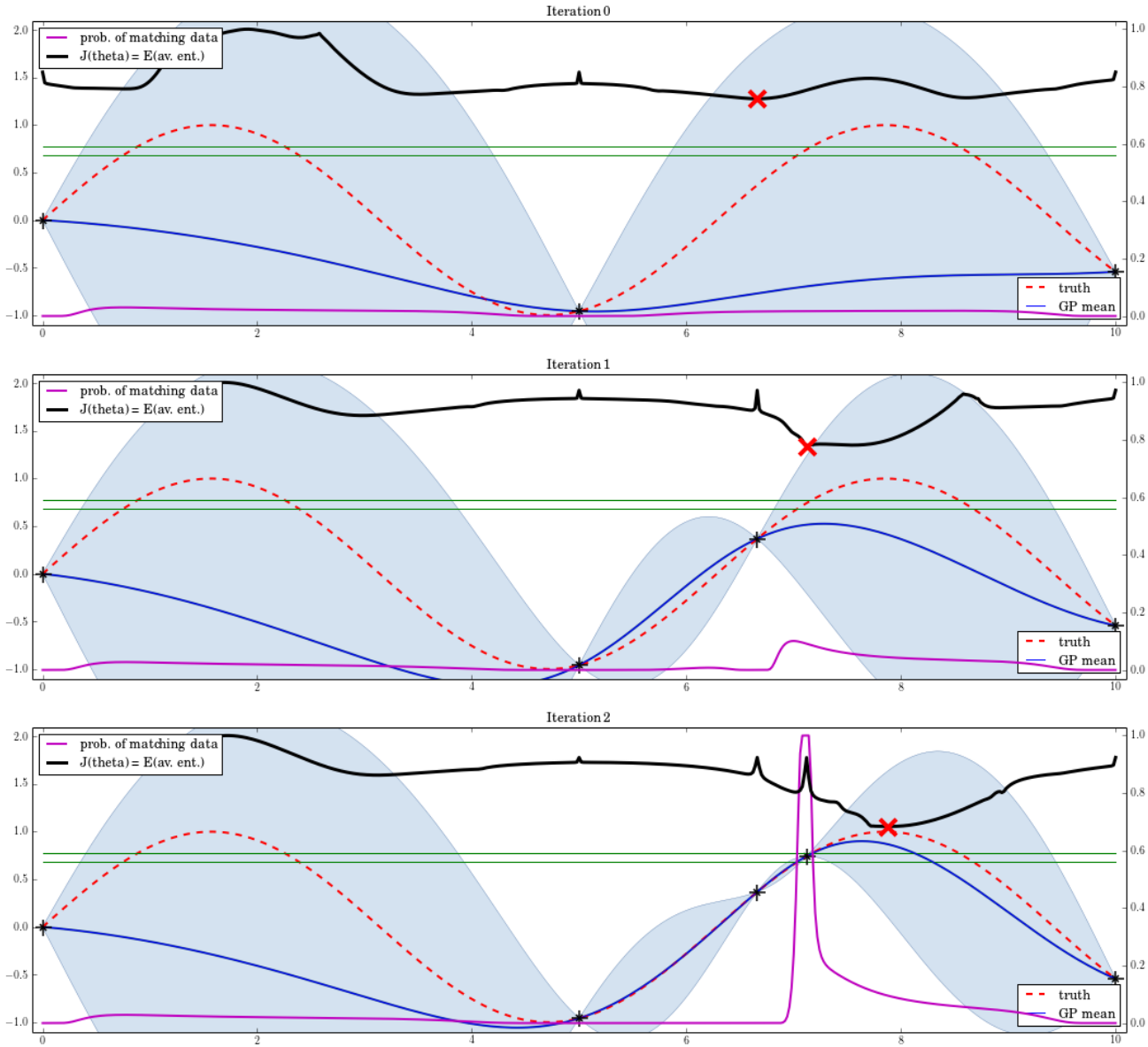
Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



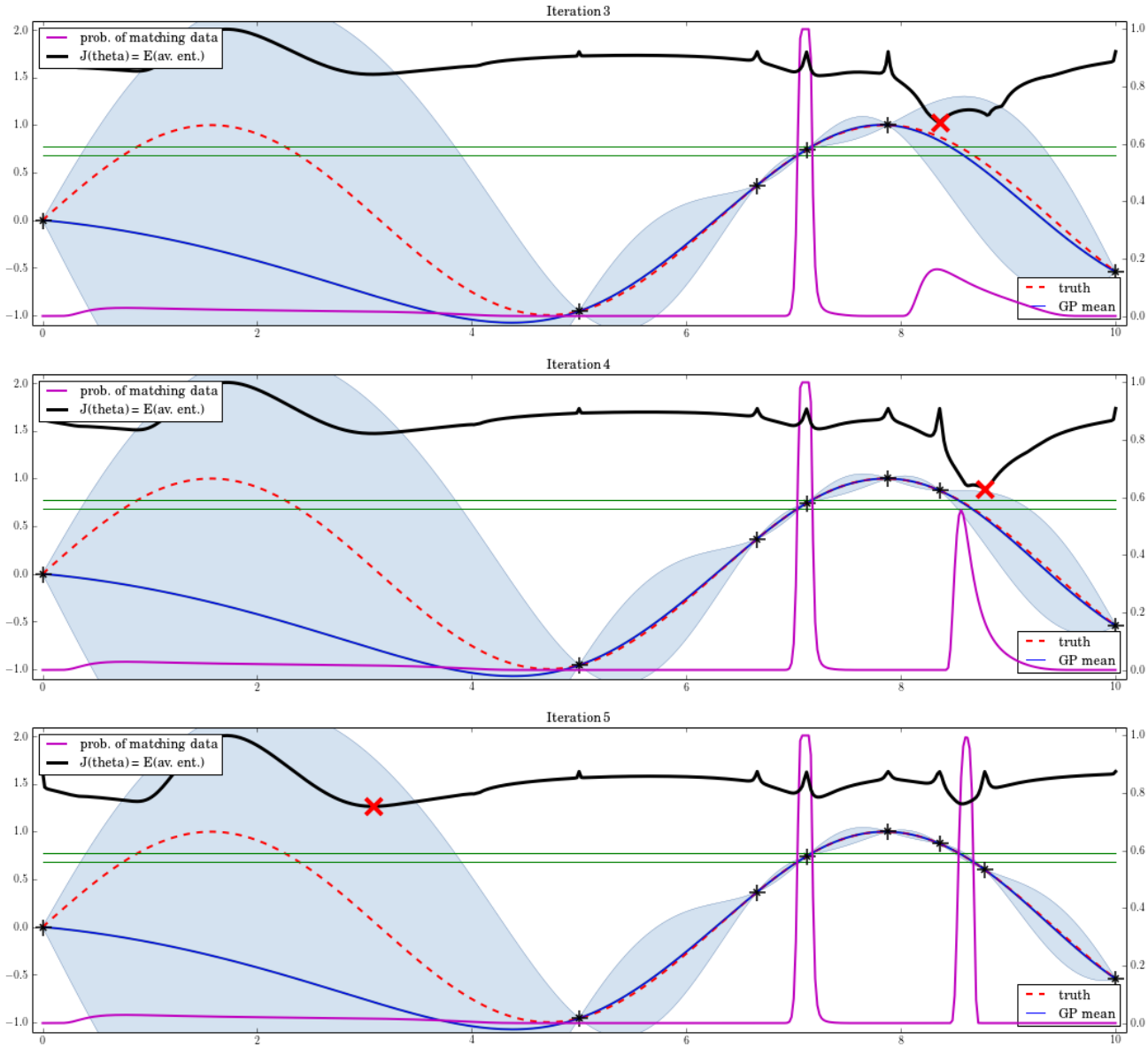
Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy

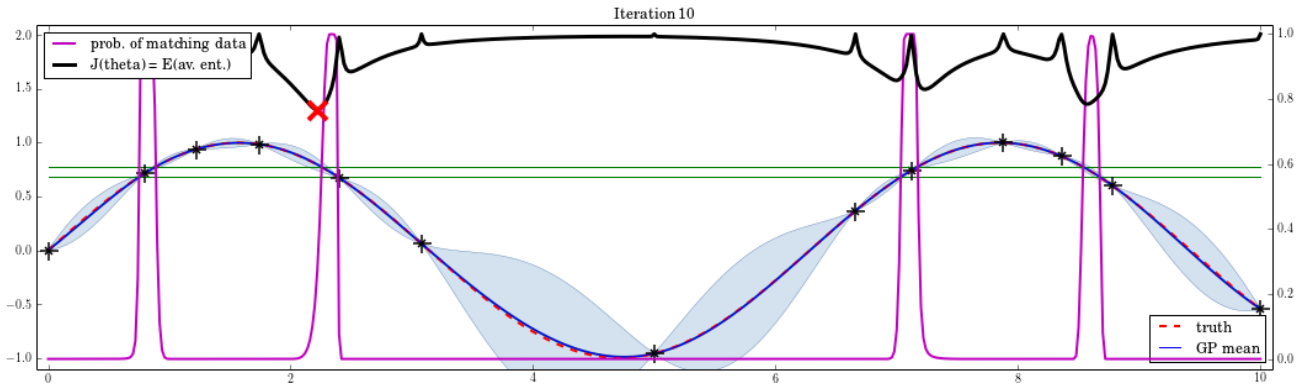


Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



Toy 1d: min expected entropy vs max entropy

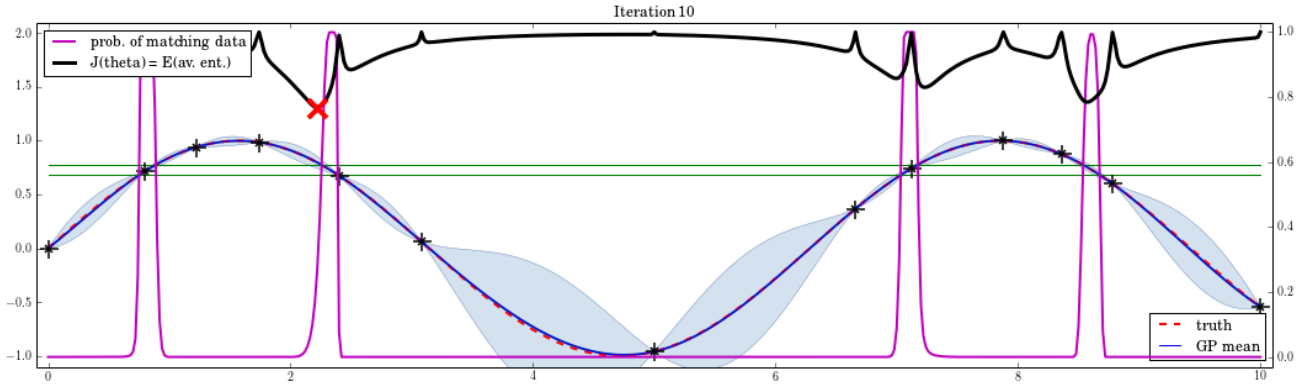
After 10 iterations, choosing the point of maximum entropy



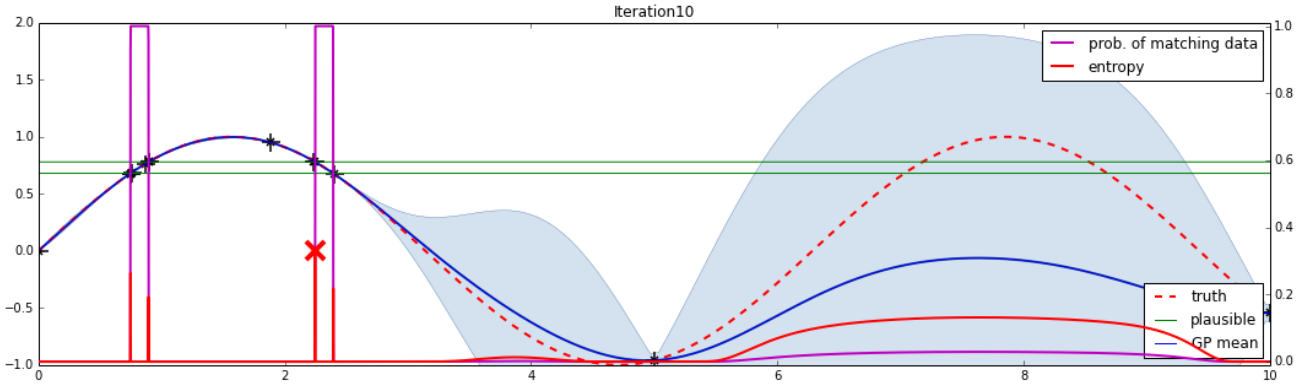
we have found the plausible region to reasonable accuracy.

Toy 1d: min expected entropy vs max entropy

After 10 iterations, choosing the point of maximum entropy



we have found the plausible region to reasonable accuracy.
Whereas maximizing the entropy has not



In 1d, a simpler space filling criterion would work just as well.

Solving the optimisation problem

Finding θ which minimises $J_n(\theta) = \mathbb{E}(E_{n+1}|\theta_{n+1} = \theta)$ is expensive.

- Even for 3d problems, grid search is prohibitively expensive
- Dynamic grids help

Solving the optimisation problem

Finding θ which minimises $J_n(\theta) = \mathbb{E}(E_{n+1}|\theta_{n+1} = \theta)$ is expensive.

- Even for 3d problems, grid search is prohibitively expensive
- Dynamic grids help

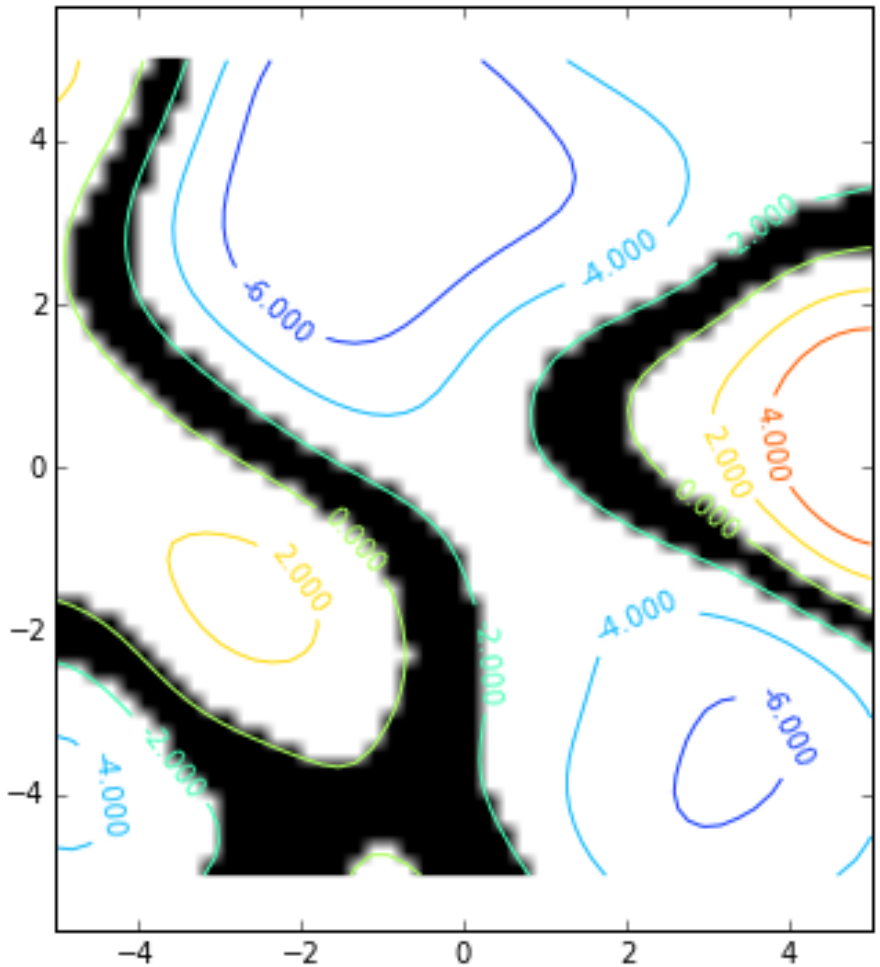
We can use Bayesian optimization to find the optima:

1. Evaluate $J_n(\theta)$ at a small number of locations
2. Build a GP model of $J_n(\cdot)$
3. Choose the next θ at which to evaluate J_n so as to minimise the expected-improvement (EI) criterion
4. Return to step 2.

History match

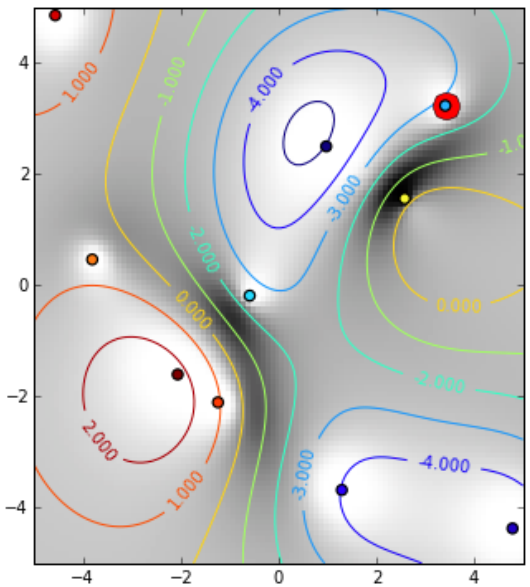
Can we learn the following plausible set?

- A sample from a GP on $\mathbb{R}^2$.
- Find x s.t. $-2 < f(x) < 0$



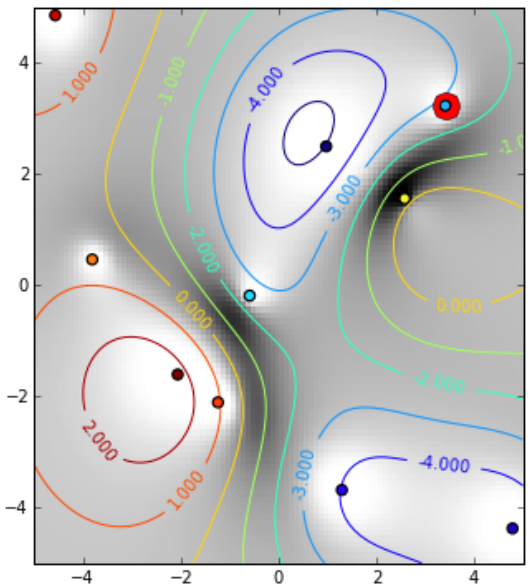
Iteration 10

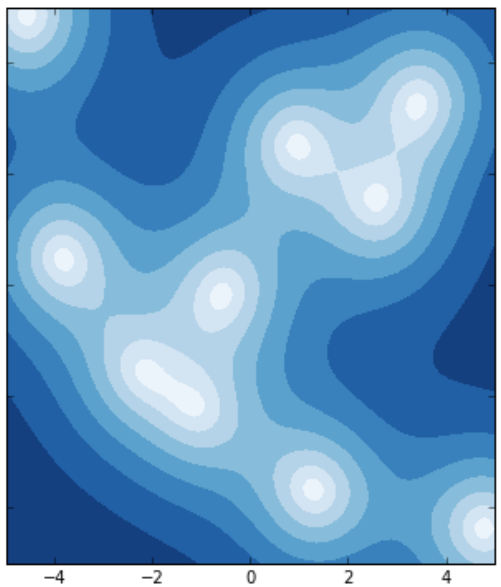
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$



Iteration 10

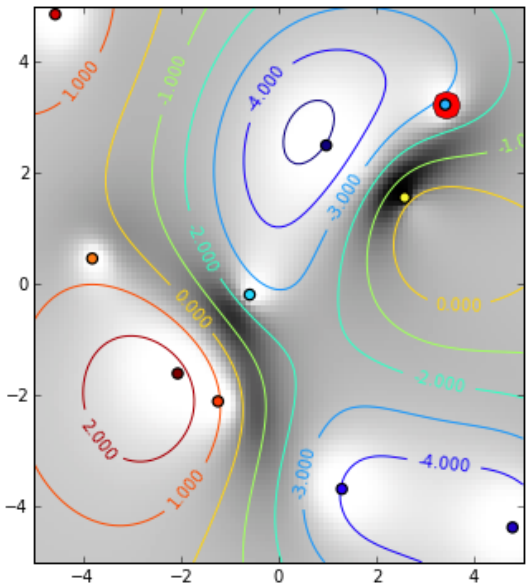
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

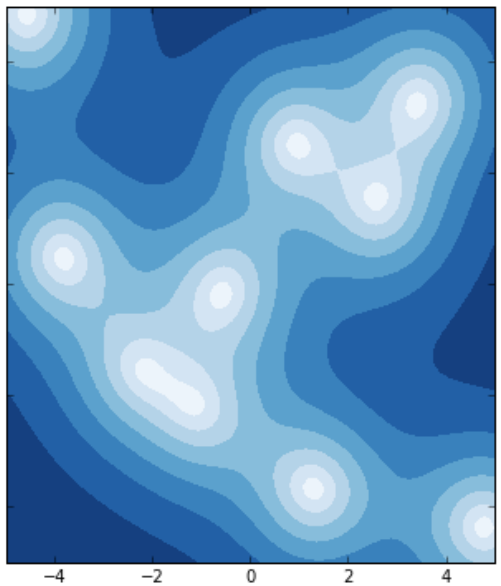


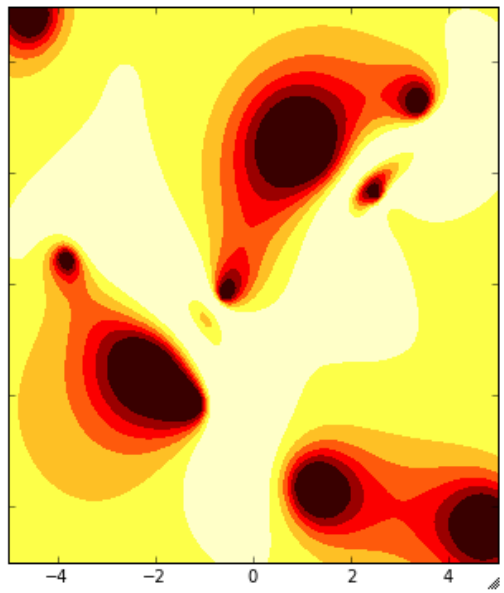


Iteration 10

Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

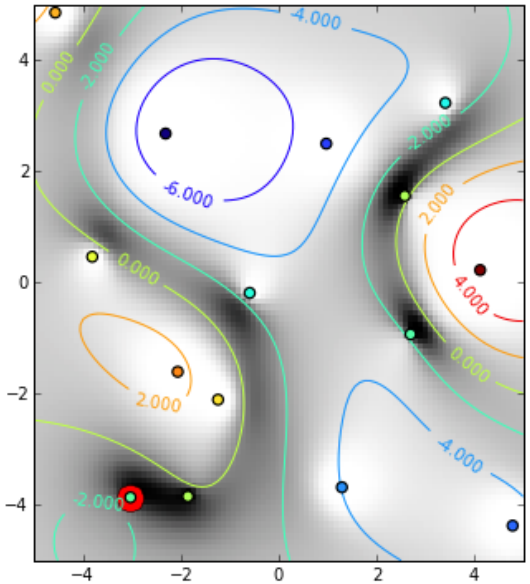


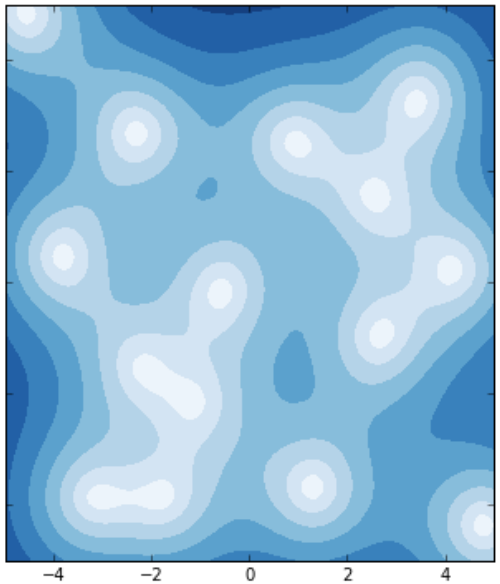


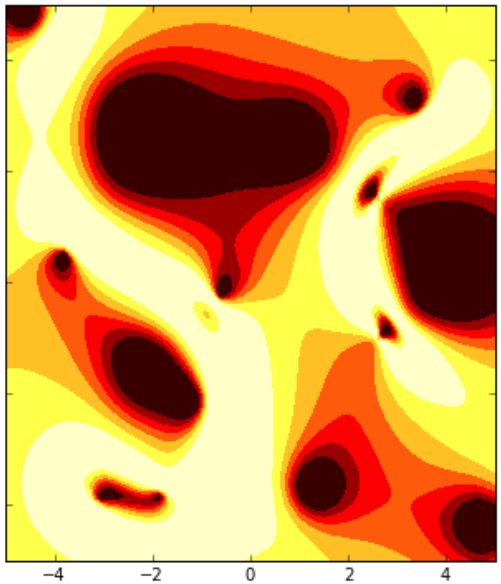


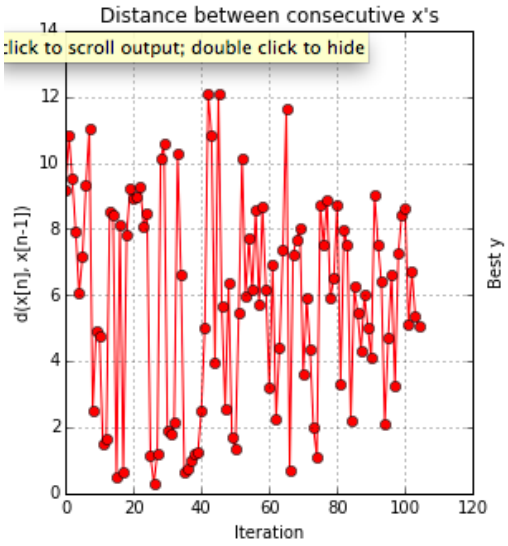
Iteration 15

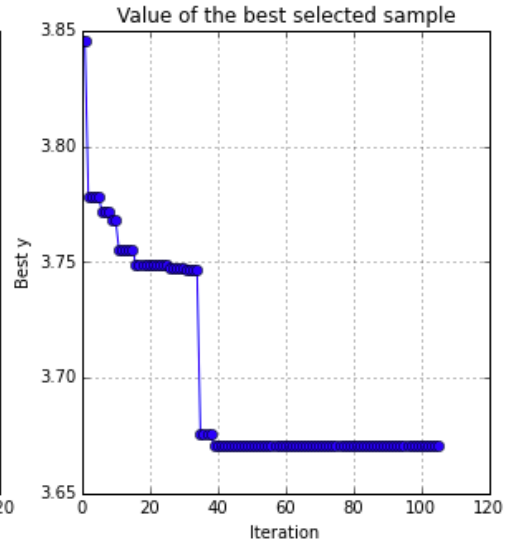
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

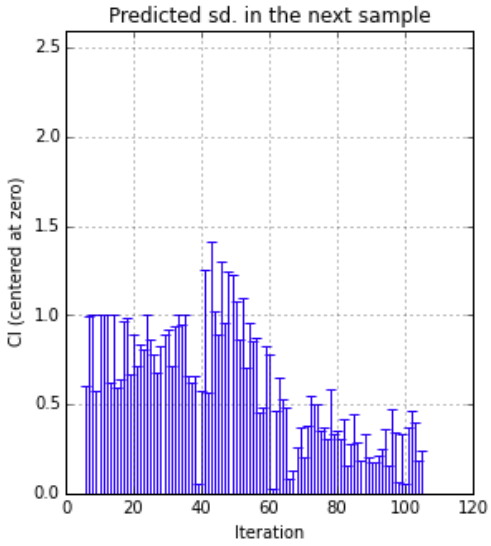




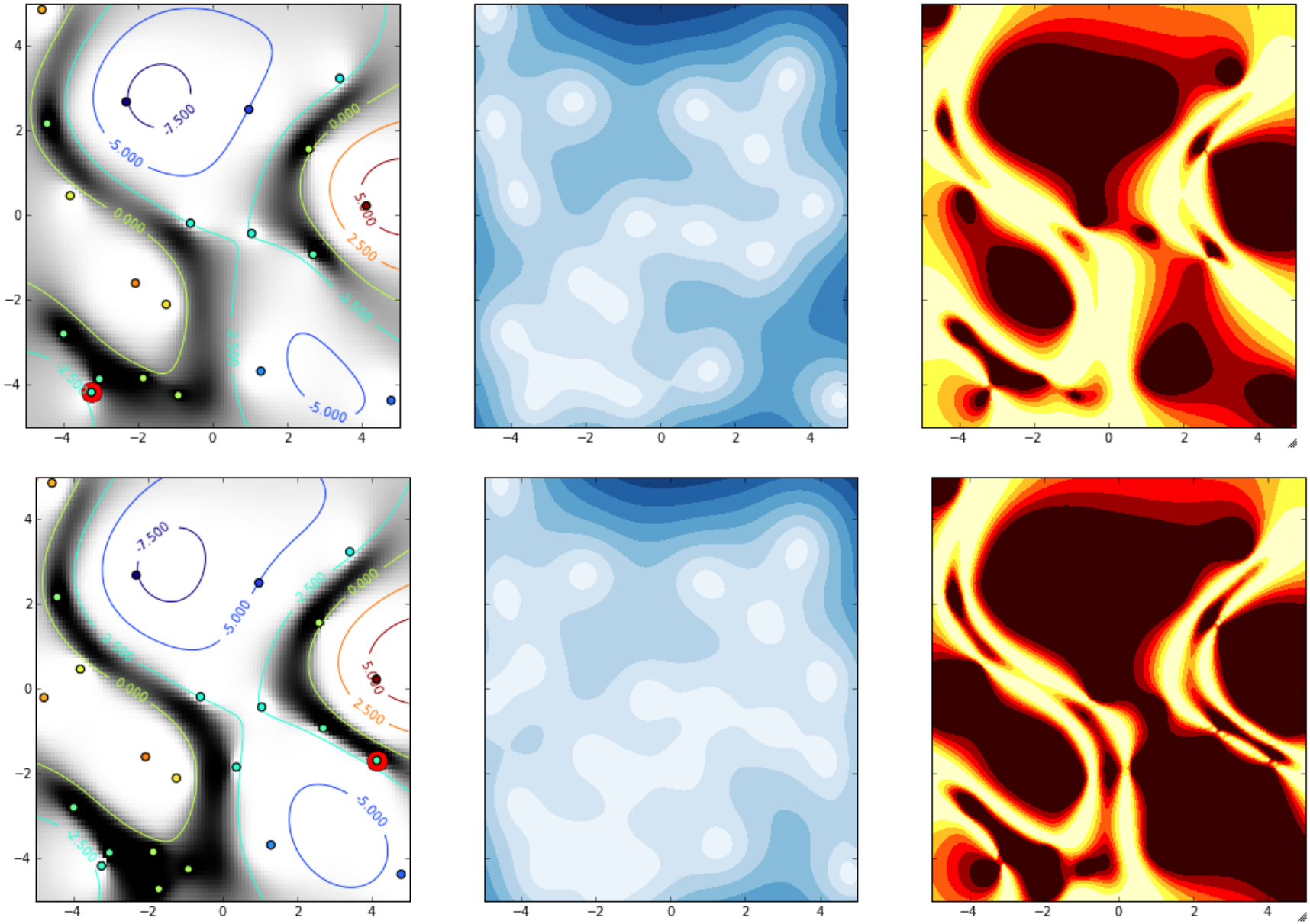








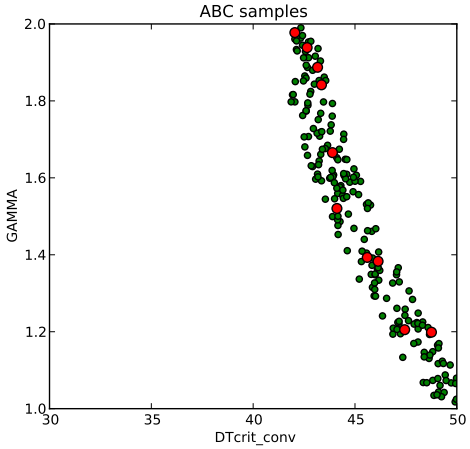
Iterations 20 and 24

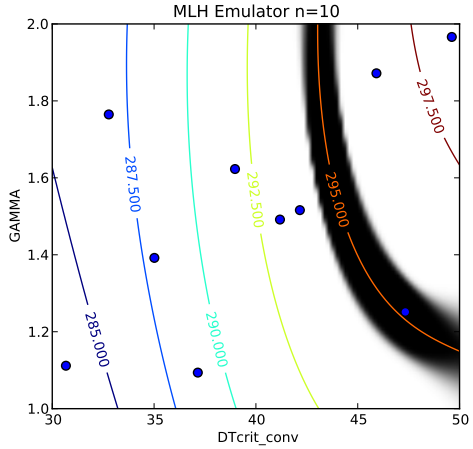


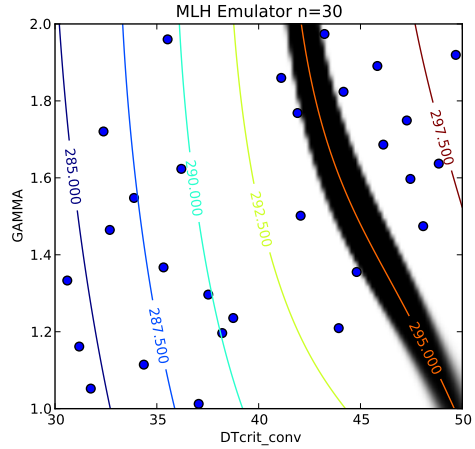
Video

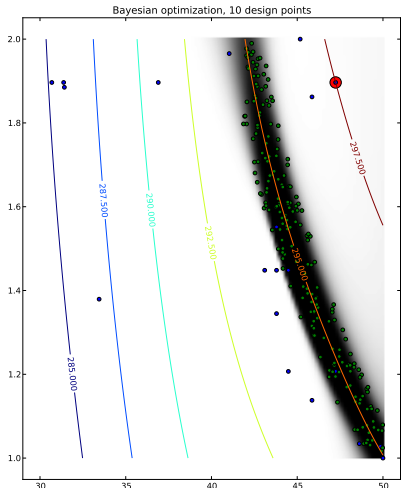
EPm: climate model

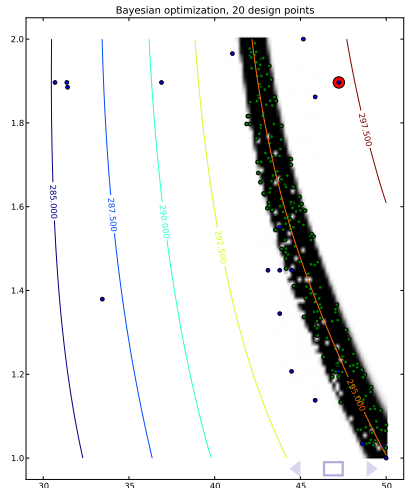
- 3d problem
- DTcrit_conv - critical temperature gradient that triggers convection
- GAMMA - emissivity parameter for water vapour
- Calibrate to global average surface temperature











Inference under discrepancy

How should we do inference if the model is imperfect?

[1] Even if we can't agree about it!

Inference under discrepancy

How should we do inference if the model is imperfect?
Data generating process

$$y \sim G$$

[1] Even if we can't agree about it!

Inference under discrepancy

How should we do inference if the model is imperfect?
Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If $G = F_{\theta_0} \in \mathcal{F}$ then we know what to do[1].

[1] Even if we can't agree about it!

Inference under discrepancy

How should we do inference if the model is imperfect?
Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If $G = F_{\theta_0} \in \mathcal{F}$ then we know what to do[1].

How should we proceed if

$$G \notin \mathcal{F}$$

[1] Even if we can't agree about it!

Inference under discrepancy

How should we do inference if the model is imperfect?
Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If $G = F_{\theta_0} \in \mathcal{F}$ then we know what to do[1].

How should we proceed if

$$G \notin \mathcal{F}$$

Interest lies in inference of θ not calibrated prediction.

[1] Even if we can't agree about it!

An appealing idea

Kennedy an O'Hagan 2001

Lets acknowledge that most models are imperfect.

An appealing idea

Kennedy an O'Hagan 2001

Lets acknowledge that most models are imperfect.

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If $f_\theta(x)$ is our simulator, y the observation, then perhaps we can correct f by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta \sim GP$$

An appealing idea

Kennedy an O'Hagan 2001

Lets acknowledge that most models are imperfect.

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If $f_\theta(x)$ is our simulator, y the observation, then perhaps we can correct f by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta \sim GP$$

This greatly expands $\mathcal{F}$ into a non-parametric world.

An appealing, but flawed, idea

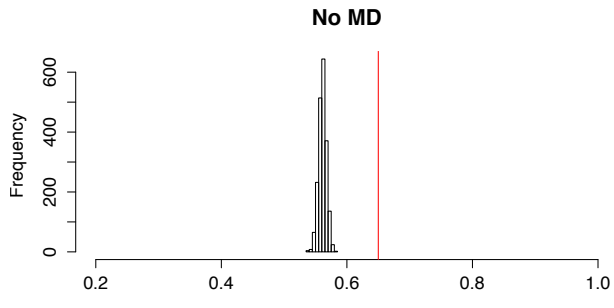
Kennedy and O'Hagan 2001, Brynjarsdottir and O'Hagan 2014

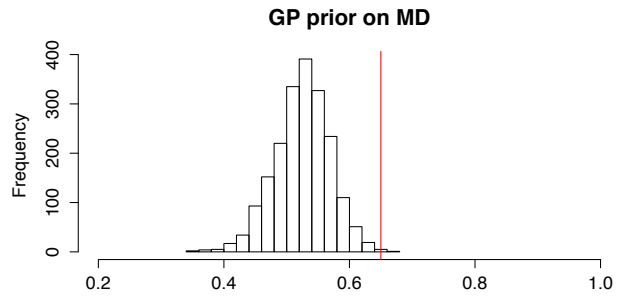
Simulator

$$f_\theta(x) = \theta x$$

Reality

$$g(x) = \frac{\theta x}{1 + \frac{x}{a}} \quad \theta = 0.65, a = 20$$





Bolting on a GP can correct your predictions, but won't necessarily fix your inference.

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is too complex to understand

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is too complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is too complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.
  - Wong et al 2017 impose identifiability (for δ and θ) by giving up and identifying

$$\theta^* = \arg\min_{\theta} \int (\zeta(x) - f_\theta(x))^2 d\pi(x)$$

J. R. Statist. Soc. B (2017)
79, *Part* 2, *pp.* 635–648

A frequentist approach to computer model calibration

Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

Anecdotally, HM and ABC seem to work better in mis-specified cases.

Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

Anecdotally, HM and ABC seem to work better in mis-specified cases.

Big question[2] is what properties would we like our inferential approach to possess?

[2] To which I have no answer

Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} l(y|\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$, then (under some conditions)

$$\hat{\theta}_n \to \theta_0 \text{ almost surely as } \quad n \to \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, \mathcal{I}^{-1}(\theta_0))$$

Asymptotic consistency, efficiency, normality.

Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} l(y|\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$, then (under some conditions)

$$\hat{\theta}_n \to \theta_0 \text{ almost surely as } \quad n \to \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, \mathcal{I}^{-1}(\theta_0))$$

Asymptotic consistency, efficiency, normality.
If $G \notin \mathcal{F}$

$$\hat{\theta}_n \to \theta^* = \arg\min_{\theta} D_{KL}(G, F_\theta) \text{ almost surely}$$

$$= \arg\min_{\theta} \int \log \frac{dG}{dF_\theta} dG$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, V^{-1})$$

Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.
This also requires (a long list of) identifiability conditions to hold.

Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.
This also requires (a long list of) identifiability conditions to hold.

If $G \notin \mathcal{F}$, we still get asymptotic concentration (and possibly normality) but to θ^* (the pseudo-true value).

> *"there is no obvious meaning for Bayesian analysis in this case"*

Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.
This also requires (a long list of) identifiability conditions to hold.

If $G \notin \mathcal{F}$, we still get asymptotic concentration (and possibly normality) but to θ^* (the pseudo-true value).

> *"there is no obvious meaning for Bayesian analysis in this case"*

Often with non-parametric models (eg GPs), we don't even get this convergence to the pseudo-true value due to lack of identifiability.

History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_{\theta,y}) \leq 3\}$$

where

$$S_{HM}(F_\theta) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\text{ar}_{F_\theta}(Y)}}$$

History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_{\theta,y}) \leq 3\}$$

where

$$S_{HM}(F_\theta) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\text{ar}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta)\mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta,y)\leq\epsilon})$$

for some choice of S (typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$) and ϵ.

History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_{\theta,y}) \leq 3\}$$

where

$$S_{HM}(F_\theta) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\text{ar}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta)\mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta,y)\leq\epsilon})$$

for some choice of S (typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$) and ϵ.

They both threshold a score, and are algorithmically comparable.

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- Asymptotic concentration or normality?

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - I wouldn't object but seems impossible for subjective priors.

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - I wouldn't object but seems impossible for subjective priors.
- Coherence?

What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - I wouldn't object but seems impossible for subjective priors.
- Coherence?
- Robustness to small mis-specifications?

Conclusions

- For complex models, surrogate-modelling approaches are often necessary
- Target of approximation: likelihood vs simulator output
  - likelihood is 1d surface, focussed on information in the data, but can be hard to model; simulator output is multi-dimensional.
- Good design can lead to substantial improvements in accuracy
  - Design needs to be specific to the task required - Space-filling designs are inefficient for calibration
  - Average entropy designs give good trade-off between exploration and defining the plausible region
- What properties do we want our inference scheme to possess?
  - If $G \notin \mathcal{F}$ can we ever hope to learn precisely about θ?
    If not we shouldn't use methods that converge/concentrate asymptotically.
- What errors should we be worrying about, and where should we spend our effort?

Conclusions

- For complex models, surrogate-modelling approaches are often necessary
- Target of approximation: likelihood vs simulator output
  - likelihood is 1d surface, focussed on information in the data, but can be hard to model; simulator output is multi-dimensional.
- Good design can lead to substantial improvements in accuracy
  - Design needs to be specific to the task required - Space-filling designs are inefficient for calibration
  - Average entropy designs give good trade-off between exploration and defining the plausible region
- What properties do we want our inference scheme to possess?
  - If $G \notin \mathcal{F}$ can we ever hope to learn precisely about θ?
    If not we shouldn't use methods that converge/concentrate asymptotically.
- What errors should we be worrying about, and where should we spend our effort?

Thank you for listening!