Gaussian process accelerated ABC

Richard Wilkinson

University of Sheffield

Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

- how do we relate simulators to reality?
- how do we estimate tunable parameters?
- how do we deal with computational constraints?

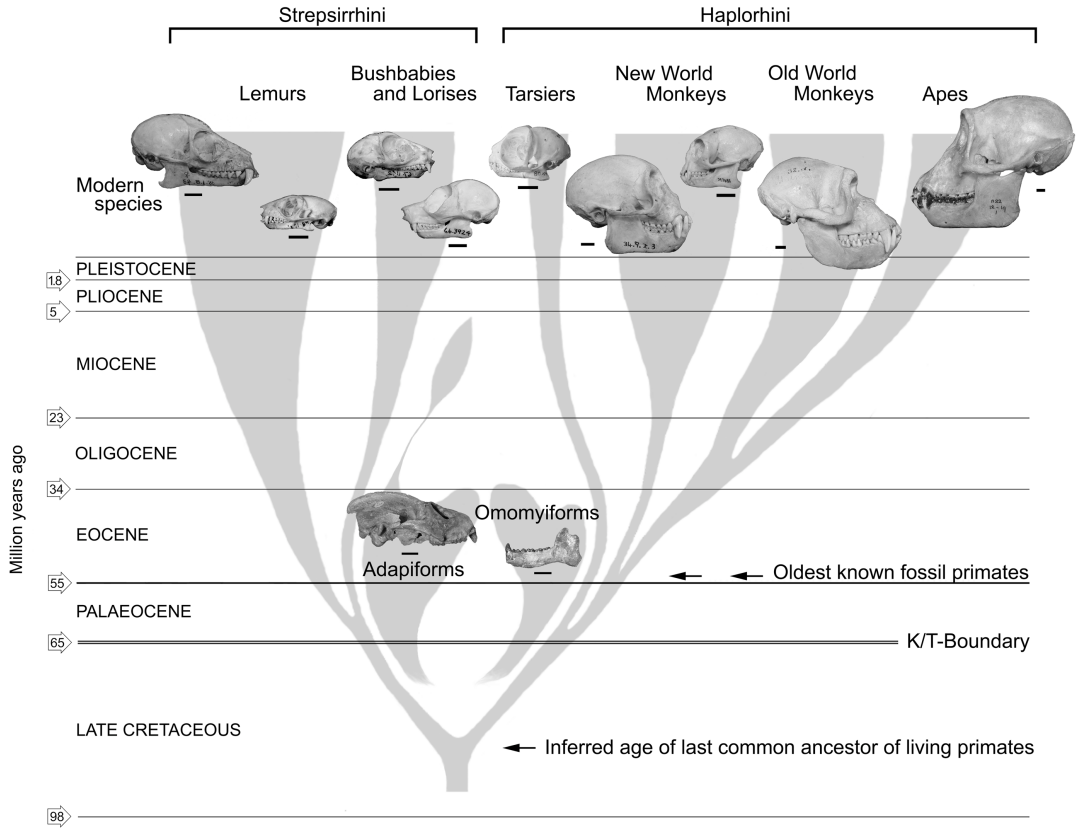
Calibration

- 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 which explain the data.

The Bayesian approach is to find the posterior distribution

$$\pi(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$



Calibration

- 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 which explain the data.

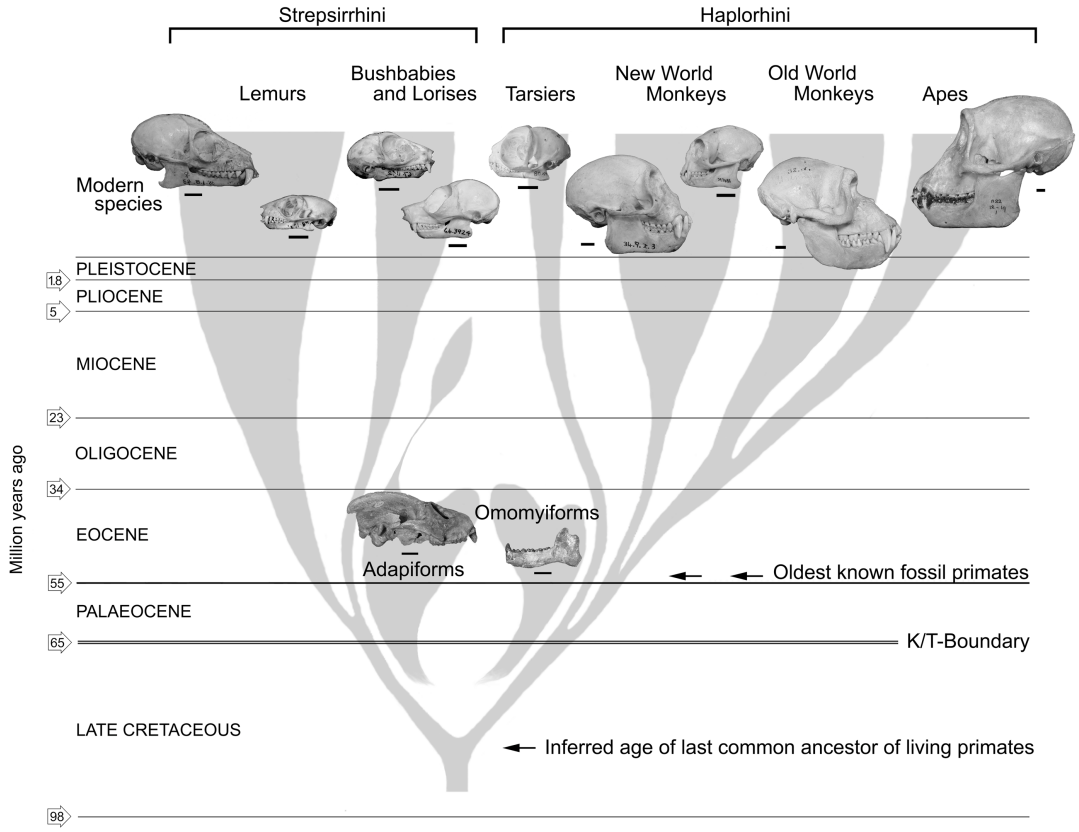
The Bayesian approach is to find the posterior distribution

$$\pi(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

For complex sims, we often don't know

$$\pi(\mathcal{D}|\theta)$$



Approximate Bayesian Computation (ABC)

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 - 'likelihood-free'

Approximate Bayesian Computation (ABC)

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 - 'likelihood-free'

ABC methods are popular in biological disciplines, particularly genetics. They are

- Simple to implement
- Intuitive
- Embarrassingly parallelizable
- Can usually be applied

'Likelihood-Free' Inference

Rejection Algorithm

- Draw θ from prior $\pi(\cdot)$
- Accept θ with probability $\pi(D \mid \theta)$

Accepted θ are independent draws from the posterior distribution, $\pi(\theta \mid D)$.

'Likelihood-Free' Inference

Rejection Algorithm

- Draw θ from prior $\pi(\cdot)$
- Accept θ with probability $\pi(D \mid \theta)$

Accepted θ are independent draws from the posterior distribution, $\pi(\theta \mid D)$.
If the likelihood, $\pi(D|\theta)$, is unknown:

'Mechanical' Rejection Algorithm

- Draw θ from $\pi(\cdot)$
- Simulate $X \sim f(\theta)$ from the computer model
- Accept θ if $D = X$, i.e., if computer output equals observation

Rejection ABC

If $\mathbb{P}(D)$ is small (or D continuous), we will rarely accept any θ. Instead, there is an approximate version:

Uniform Rejection Algorithm

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

Rejection ABC

If $\mathbb{P}(D)$ is small (or D continuous), we will rarely accept any θ. Instead, there is an approximate version:

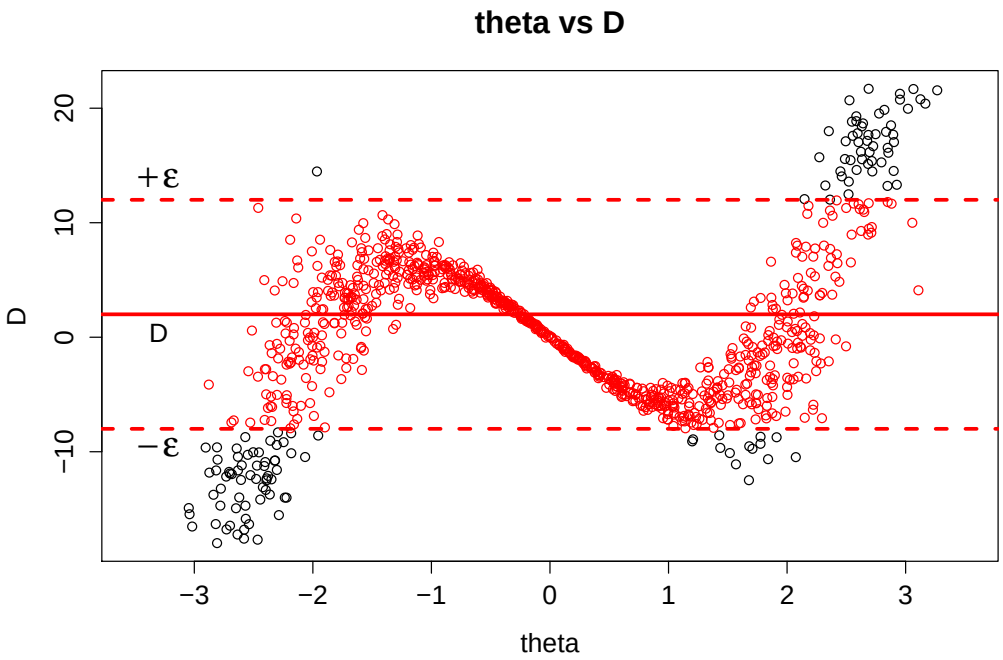
Uniform Rejection Algorithm

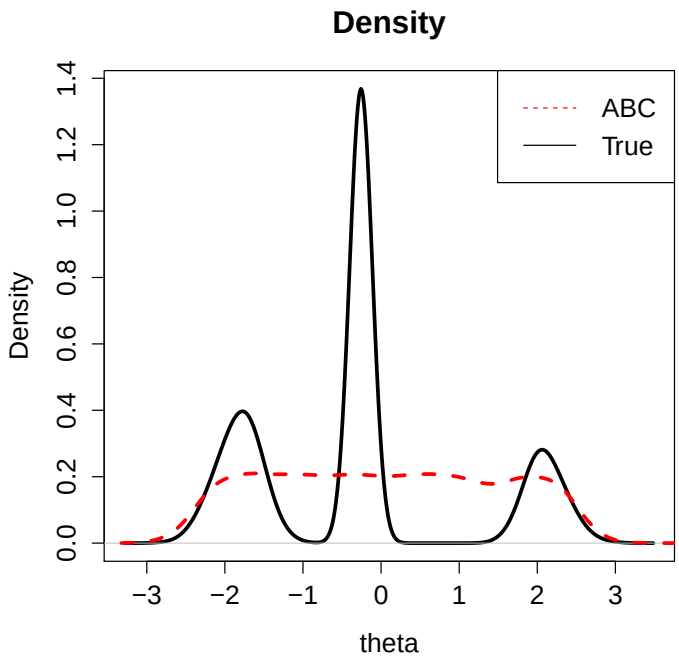
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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 $\pi(\theta \mid D)$.

$\epsilon = 10$

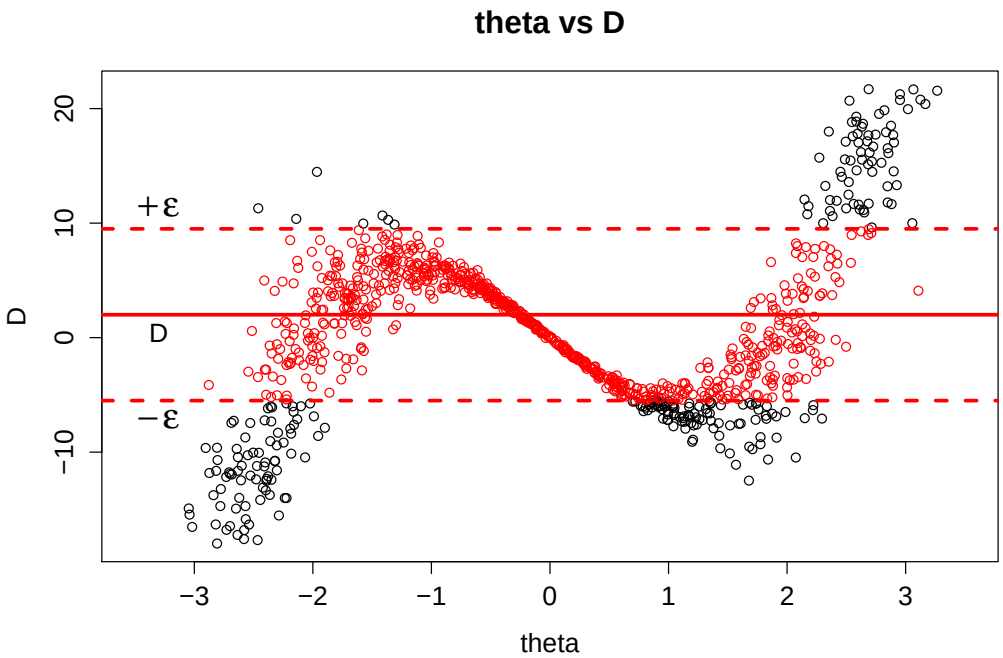


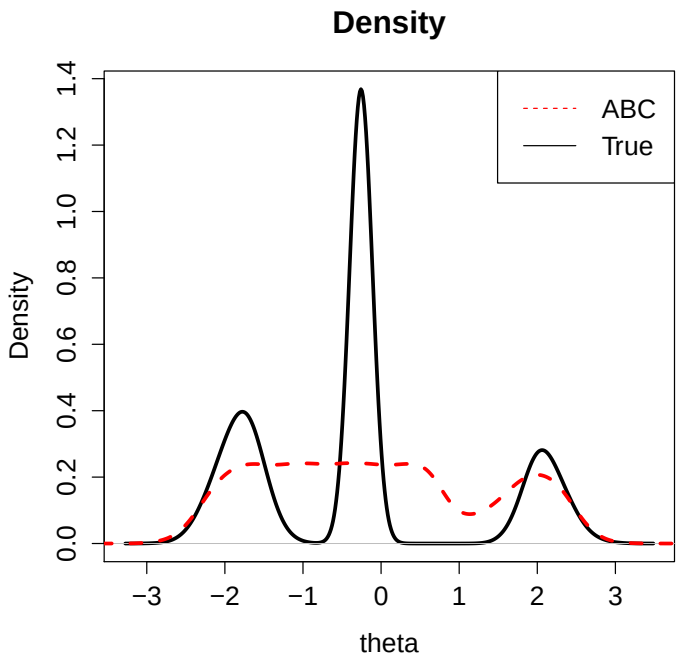


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

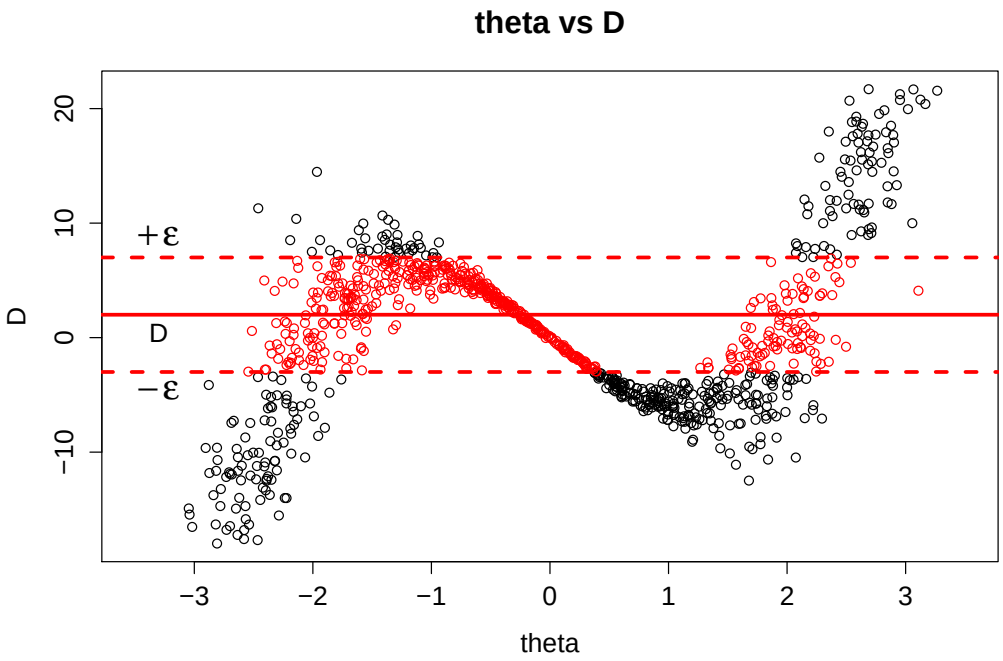
$$\rho(D, X) = |D - X|, \qquad D = 2$$

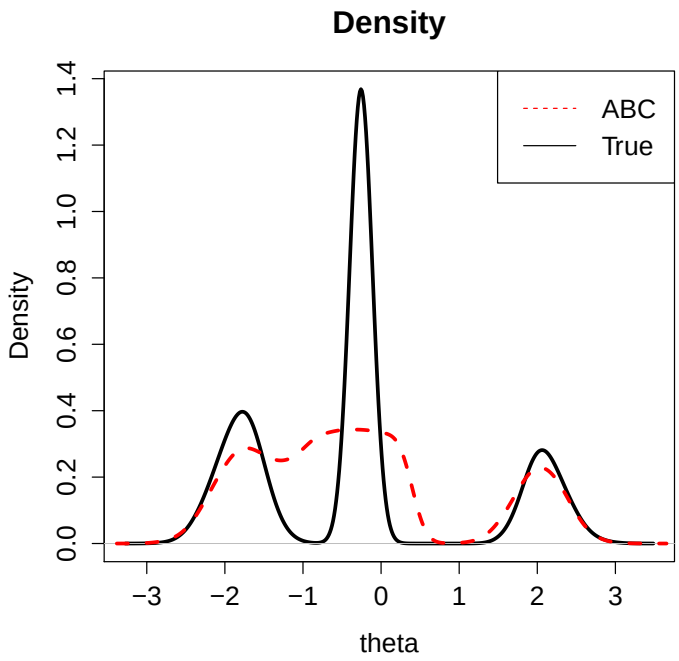
$\epsilon = 7.5$


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

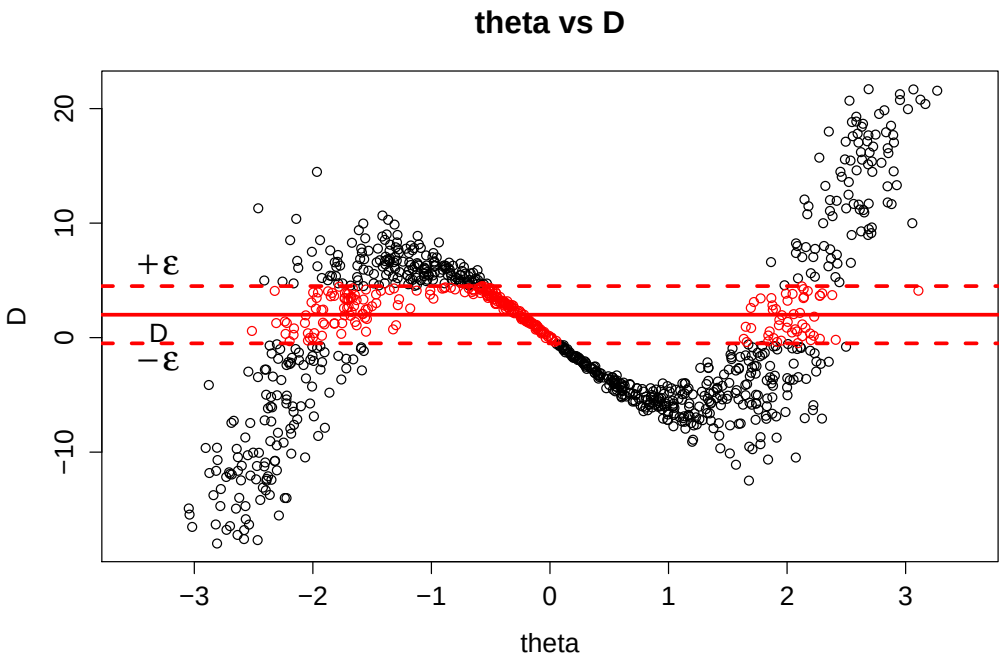

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

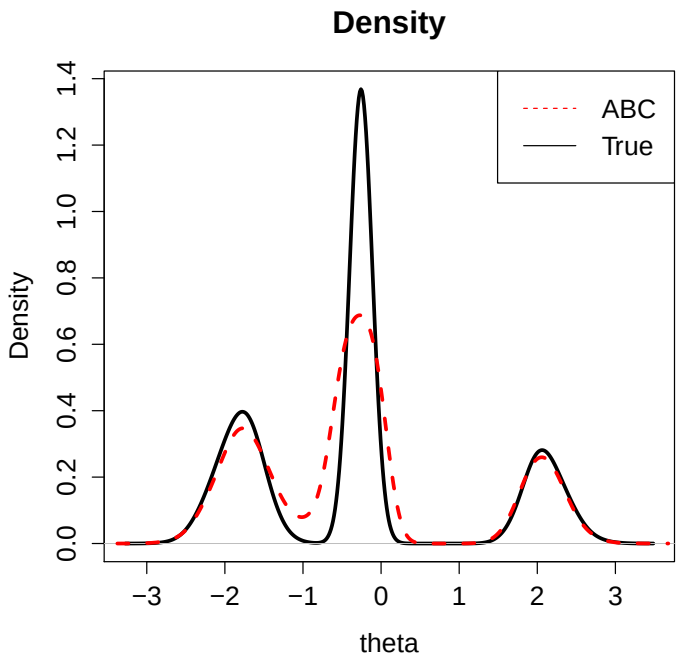
$\epsilon = 5$


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

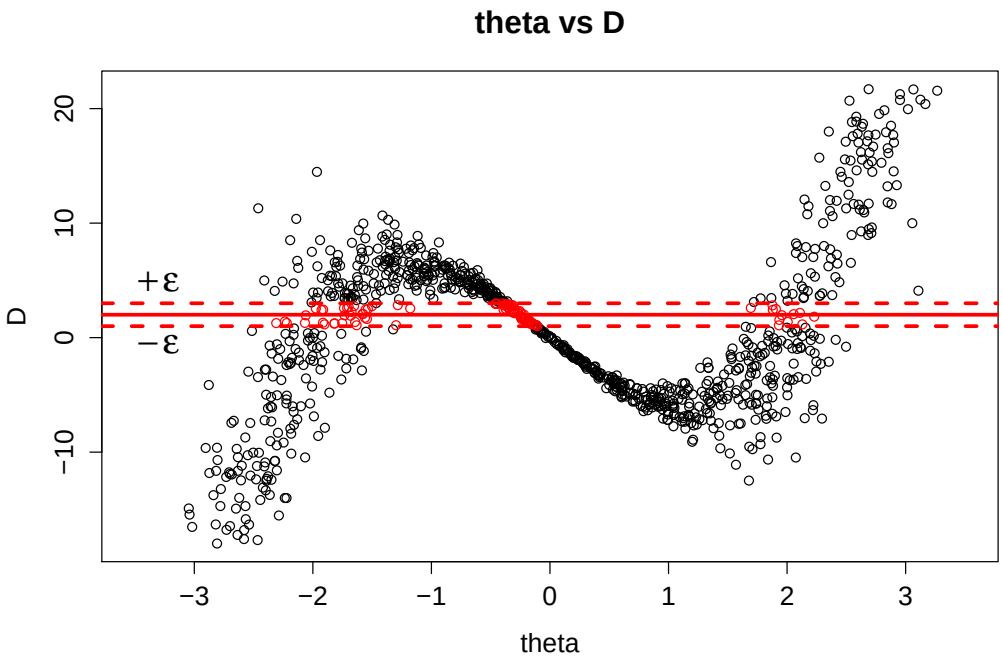

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

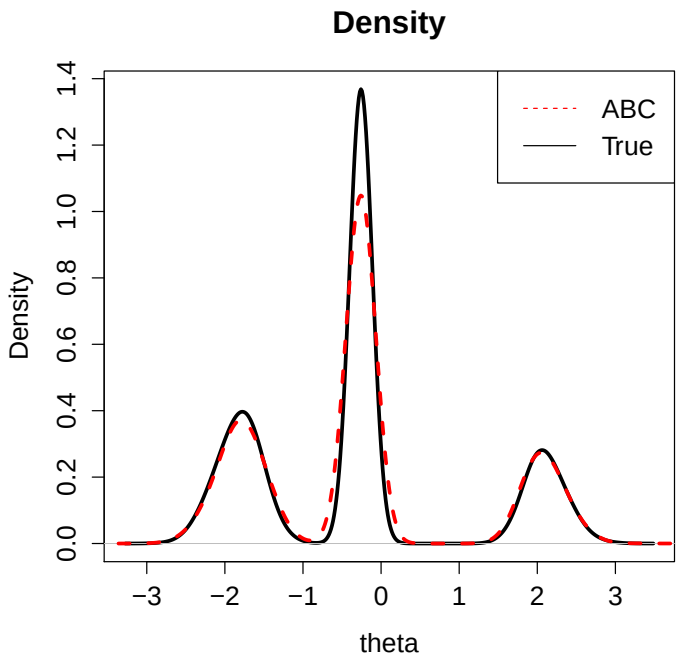
$\epsilon = 2.5$


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

Rejection ABC

If the data are too high dimensional we never observe simulations that are 'close' to the data

Reduce the dimension using summary statistics, $S(D)$.

Approximate Rejection Algorithm With Summaries

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

If S is sufficient this is equivalent to the previous algorithm.

Rejection ABC

If the data are too high dimensional we never observe simulations that are 'close' to the data

Reduce the dimension using summary statistics, $S(D)$.

Approximate Rejection Algorithm With Summaries

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

If S is sufficient this is equivalent to the previous algorithm.

Simple $\rightarrow$ Popular with non-statisticians

Expensive stochastic simulators

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.
- each simulation runs MCMC of Hamiltonian dynamics
- want to infer number of stem cells by comparing patterns with real data
- Each simulation takes about an hour, and is stochastic.

Efficient algorithms can take us only so far...

We will continue face situations in which we are limited by computer power.

Bayesian inference for computer experiments

Emulation/surrogate modelling/meta-modelling

Sacks *et al.* 1989 introduce the idea of an *emulator*

- if $f(x)$ is an expensive (deterministic) simulator, approximate it by a cheaper surrogate model

Kennedy and O'Hagan 2001 used emulators to solve a Bayesian inference problem

Bayesian calibration of **computer models**
MC Kennedy, A O'Hagan - Journal of the Royal Statistical ..., 2001 - Wiley Online Library
Summary. We consider prediction and uncertainty analysis for systems which are approximated using complex mathematical models. Such models, implemented as computer codes, are often generic in the sense that by a suitable choice of some of the model's input ...
Cited by 1587 Related articles All 22 versions Web of Science: 759 Cite Save More

Bayesian inference for computer experiments

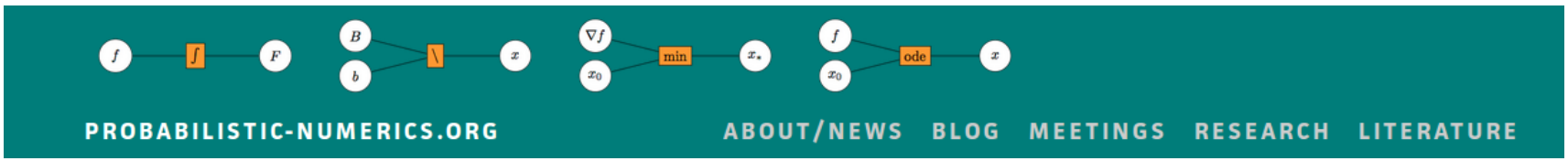
Emulation/surrogate modelling/meta-modelling

Sacks *et al.* 1989 introduce the idea of an *emulator*

- if $f(x)$ is an expensive (deterministic) simulator, approximate it by a cheaper surrogate model

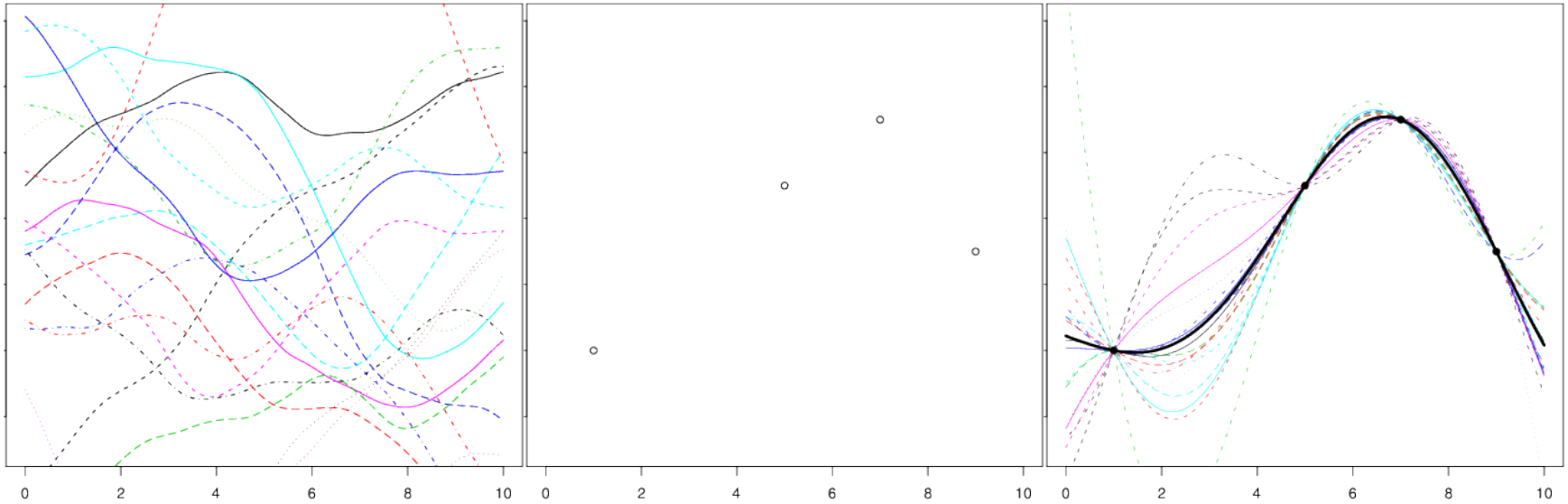
Kennedy and O'Hagan 2001 used emulators to solve a Bayesian inference problem

Bayesian calibration of **computer models**
MC Kennedy, A O'Hagan - Journal of the Royal Statistical ..., 2001 - Wiley Online Library
Summary. We consider prediction and uncertainty analysis for systems which are approximated using complex mathematical models. Such models, implemented as computer codes, are often generic in the sense that by a suitable choice of some of the model's input ...
Cited by 1587 Related articles All 22 versions Web of Science: 759 Cite Save More



Numerical algorithms, e.g. integration, solving O/PDEs, optimization, estimate some unknown quantity on the basis of function evaluations, ie, they are inference problems.

If in doubt, use a Gaussian process



- Bayesian quadrature: Diaconis 1988, O'Hagan 1991,

$$\int f(x)dx$$

  Replace f by a GP - the integral is then Gaussian.
- Bayesian optimization: find $\arg\max f(x)$ with a minimum number of function calls.
- ...

Synthetic likelihood

Wood 2010 introduced synthetic likelihood:

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

where μ_θ and Σ_θ are the mean and covariance of $X_1, \ldots, X_n$ which are outputs from the simulator run at θ. This is then plugged into an MCMC sampler.

Synthetic likelihood

Wood 2010 introduced synthetic likelihood:

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

where μ_θ and Σ_θ are the mean and covariance of $X_1, \ldots, X_n$ which are outputs from the simulator run at θ. This is then plugged 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*

Surrogate-model ABC papers

- Wilkinson 2014
- Meeds and Welling 2014
- **Gutmann and Corander 2015**
- +Others

GP-ABC

Constituent elements:

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

Target of approximation

What should we approximate with the surrogate model?

- Simulator output (Kennedy and O'Hagan 2001, Henderson et al. 2009, Meeds and Welling, 2014), for example, within a synthetic likelihood approach

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

$L(\theta) = N(D; \mu_\theta, \Sigma_\theta)$ and model

$$\mu_\theta \sim GP(\cdot, \cdot) \qquad \Sigma_\theta \sim GP(\cdot, \cdot)$$

Target of approximation

What should we approximate with the surrogate model?

- Simulator output (Kennedy and O'Hagan 2001, Henderson et al. 2009, Meeds and Welling, 2014), for example, within a synthetic likelihood approach

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

$L(\theta) = N(D; \mu_\theta, \Sigma_\theta)$ and model

$$\mu_\theta \sim GP(\cdot, \cdot) \qquad \Sigma_\theta \sim GP(\cdot, \cdot)$$

  - often easy to work with
  - hard if $S(X)$ is high dimensional
  - Often assume $\Sigma_\theta = \text{diag}(\Sigma_\theta)$ and build independent surrogates
  - requires a global approximation, i.e., need to predict $f(\theta)$ at all θ of interest.
  - Gaussian likelihood (either of the GP or the synthetic likelihood) often a poor choice for stochastic simulators

Target of approximation

What should we approximate with the surrogate model?

- (ABC) Likelihood function (Wilkinson 2014), for example

$$L_{ABC}(\theta) = \mathbb{E}_{X|\theta} K_\epsilon[\rho(S(D), S(X))] \equiv \mathbb{E}_{X|\theta} \pi_\epsilon(D|X)$$

or

$$L_{SL}(\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$

Target of approximation

What should we approximate with the surrogate model?

- (ABC) Likelihood function (Wilkinson 2014), for example

$$L_{ABC}(\theta) = \mathbb{E}_{X|\theta} K_\epsilon[\rho(S(D), S(X))] \equiv \mathbb{E}_{X|\theta} \pi_\epsilon(D|X)$$

or

$$L_{SL}(\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$

Target of approximation

What should we approximate with the surrogate model?

- (ABC) Likelihood function (Wilkinson 2014), for example

$$L_{ABC}(\theta) = \mathbb{E}_{X|\theta} K_\epsilon[\rho(S(D), S(X))] \equiv \mathbb{E}_{X|\theta} \pi_\epsilon(D|X)$$

or

$$L_{SL}(\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$$

  - 1 dimensional output surface
  - allows us to focus on the data, i.e., predict $\log L(\theta)$ at all θ. The data D is fixed
  - hard to model
  - hard to gain physical insights - primarily useful for calibration

Target of approximation

What should we approximate with the surrogate model?

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

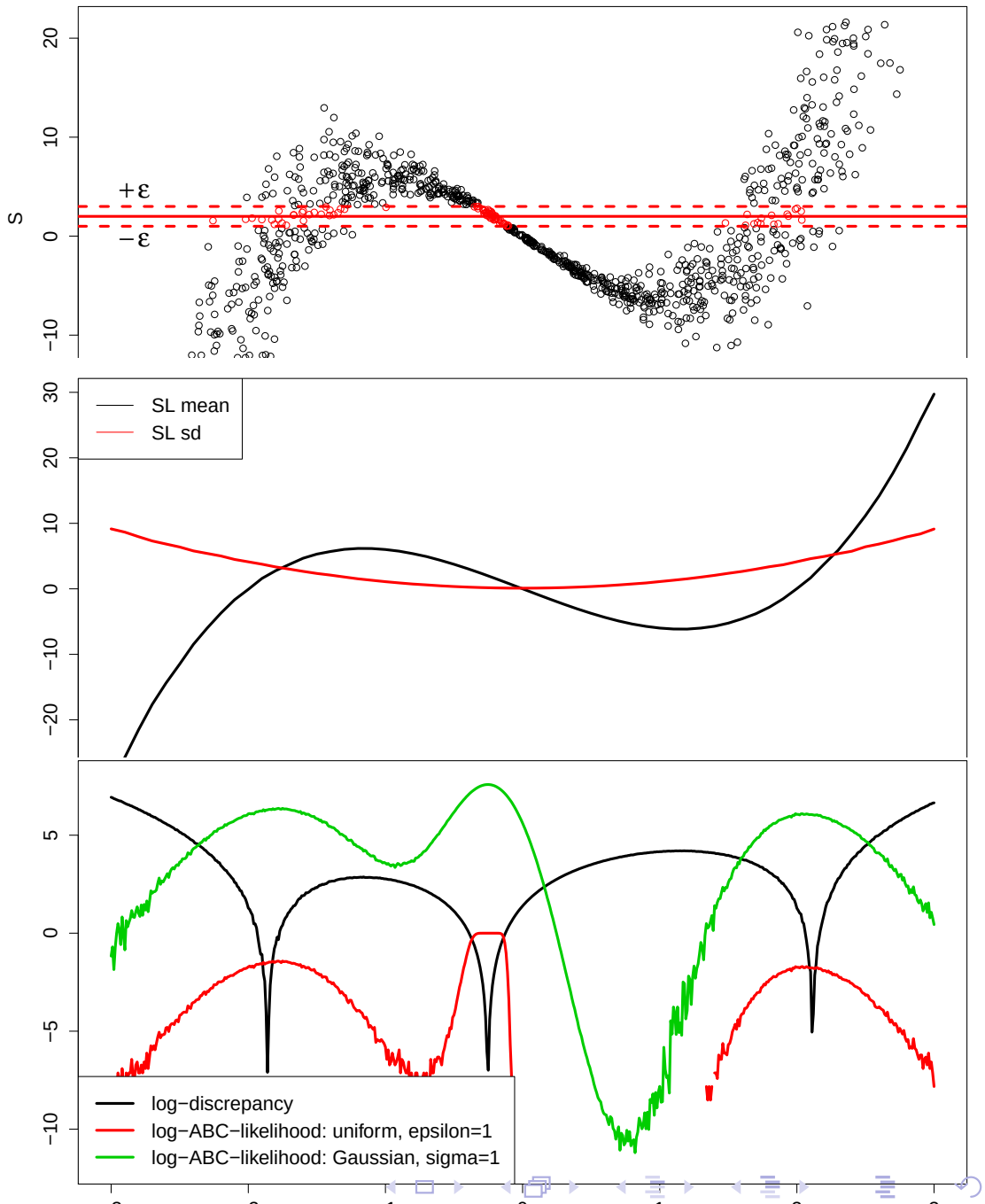
$$J(\theta) = \mathbb{E}\rho(S(D), S(X))$$

  - Also 1d, and focused on data
  - Doesn't depend upon kernel, bandwidth/tolerance etc
  - Lack of interpretability of output distributions - lose any statistical model interpretation
  - No longer targeting a posterior distribution - what are we doing?

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

Synthetic likelihood:

ABC likelihood and discrepancy:



Aim of the inference

Probabilistic calibration
Find the posterior distribution

$$\pi_{ABC}(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

for likelihood function

$$\pi_{ABC}(\mathcal{D}|\theta) = \int \pi_\epsilon(D|X)\pi(X|\theta)\mathrm{d}X$$

History matching
Find the plausible parameter set

$$\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$$

where $\mathcal{P}_D$ is some plausible set of simulation outcomes consistent with the data and errors

$$\mathcal{P}_D = \{X : |D - X| \leq 3(\sigma_e + \sigma_\epsilon)\}$$

Aim of the inference

Probabilistic calibration
Find the posterior distribution

$$\pi_{ABC}(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

for likelihood function

$$\pi_{ABC}(\mathcal{D}|\theta) = \int \pi_\epsilon(D|X)\pi(X|\theta)\mathrm{d}X$$

History matching
Find the plausible parameter set

$$\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$$

where $\mathcal{P}_D$ is some plausible set of simulation outcomes consistent with the data and errors

$$\mathcal{P}_D = \{X : |D - X| \leq 3(\sigma_e + \sigma_\epsilon)\}$$

Calibration finds a distribution representing plausible parameter values;
History matching classifies parameter space as plausible or implausible.
Other approaches such as Gutmann and Corander 2015 minimize the discrepancy to find good parameters, with less(?) of a focus on uncertainty.

History matching waves

Wilkinson 2014

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

History matching waves

Wilkinson 2014

The ABC log-likelihood $l(\theta) = \log L(\theta)$ typical ranges across a wide 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.

History matching waves

Wilkinson 2014

The ABC log-likelihood $l(\theta) = \log L(\theta)$ typical ranges across a wide 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.

Choose threshold T so that if $l(\hat{\theta}) - l(\theta) > T$ then $\pi(\theta|y) \approx 0$.

We decide that θ is implausible if

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

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

- Ruling θ to be implausible is to set $\pi(\theta|y) = 0$
- Equivalent to doing inference with log-likelihood $L(\theta)\mathbb{I}_{l(\hat{\theta})-l(\theta)<T}$

Choice of T is problem specific; start conservatively with T large and decrease, cf sequential ABC.

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.

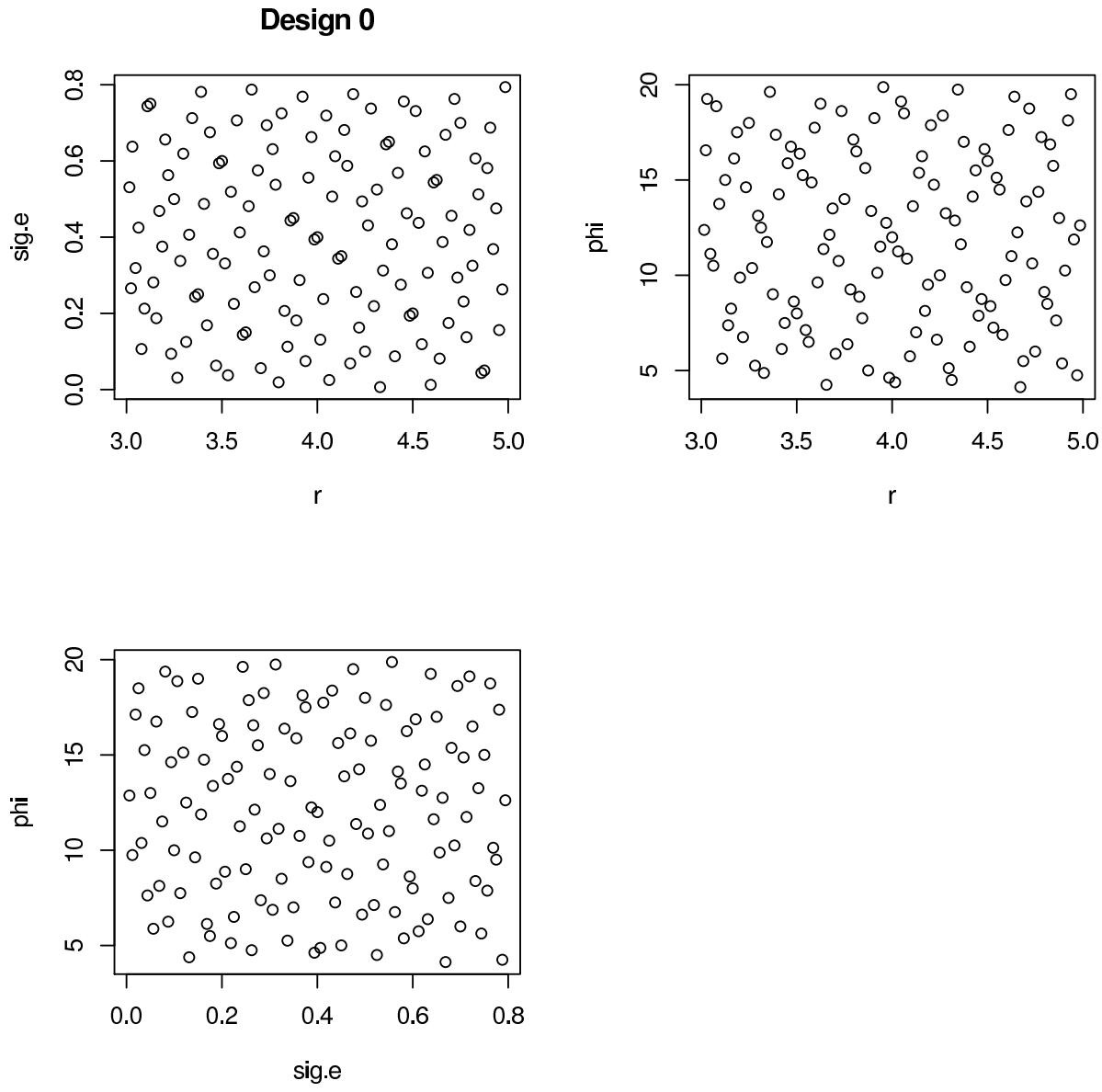
$$N_{t+1} = rN_t e^{-N_t + e_r}$$

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

- Assume we observe counts y_t where

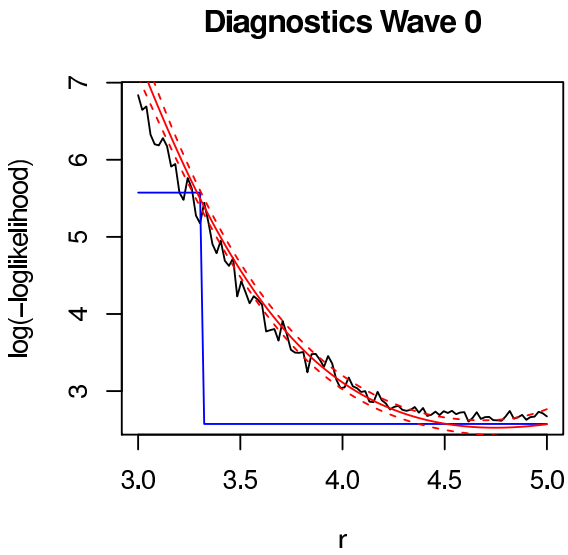
$$y_t \sim Po(\phi N_t)$$

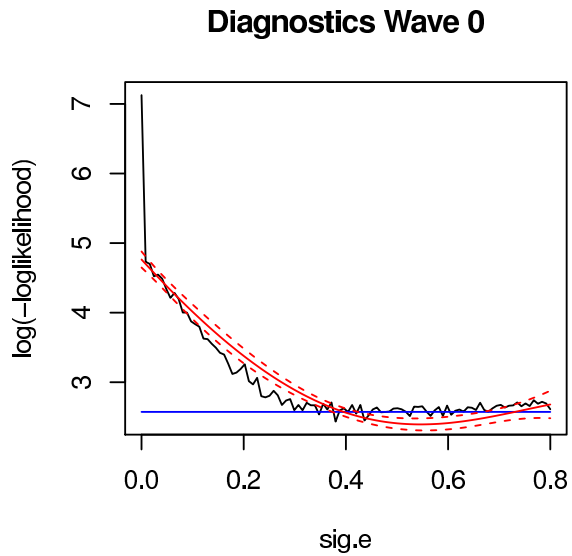
Results - Design 1 - 128 pts

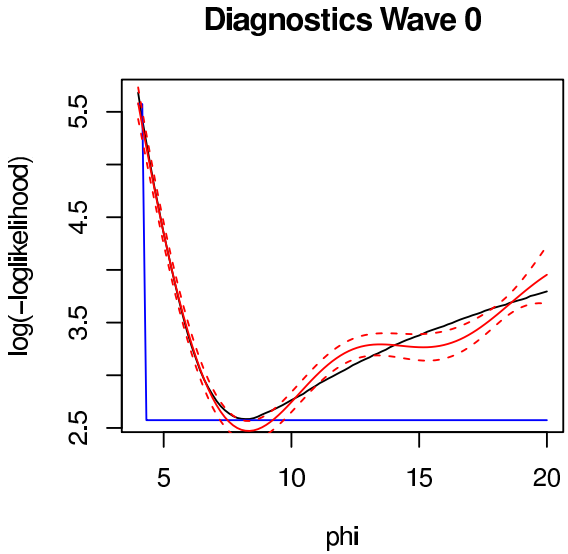


Diagnostics for GP 1 modelling $\log(-\log l(\theta))$

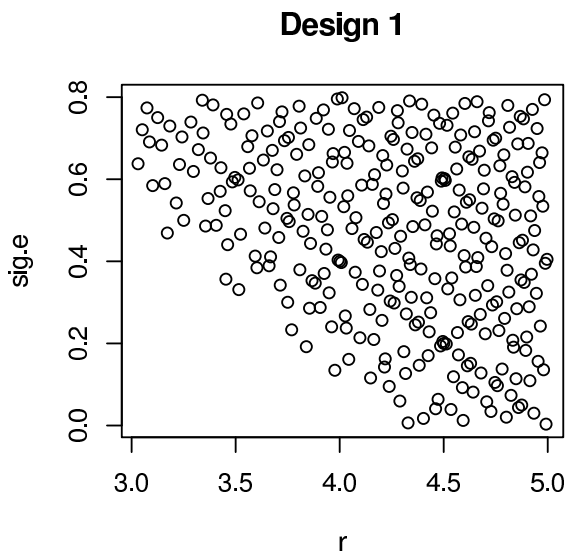
Threshold = 5.6

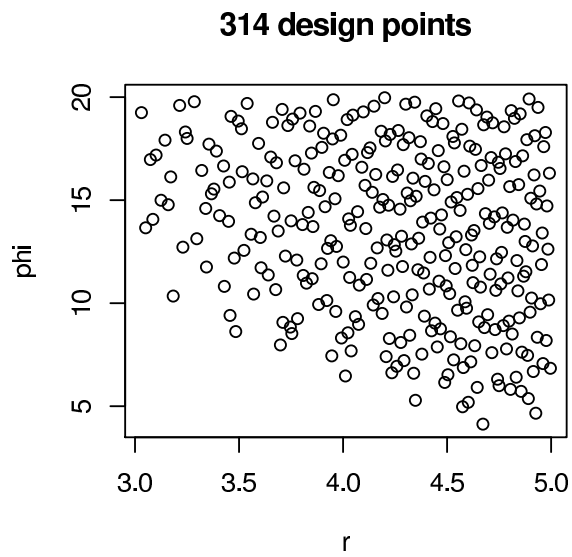


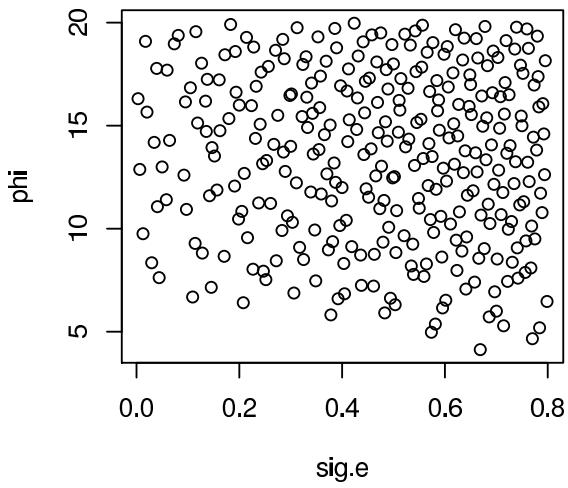




Results - Design 2 - 314 pts - 38% of space implausible

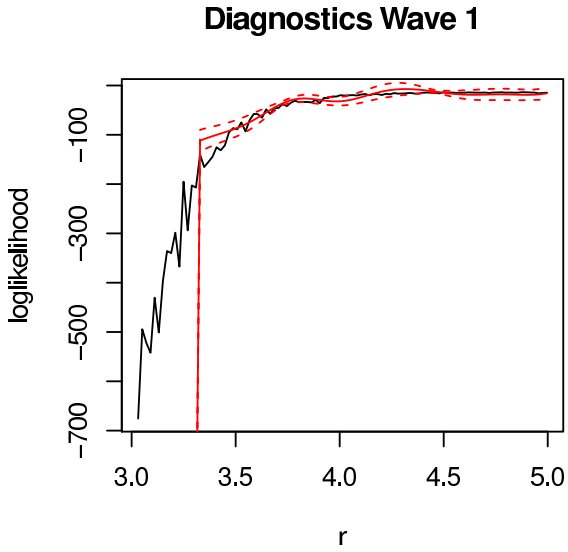


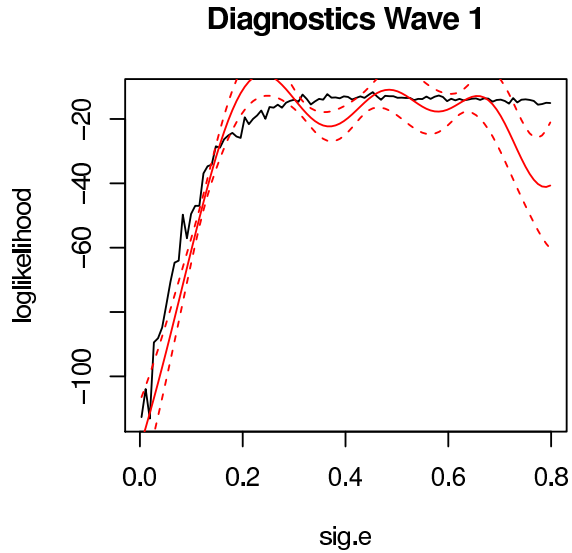


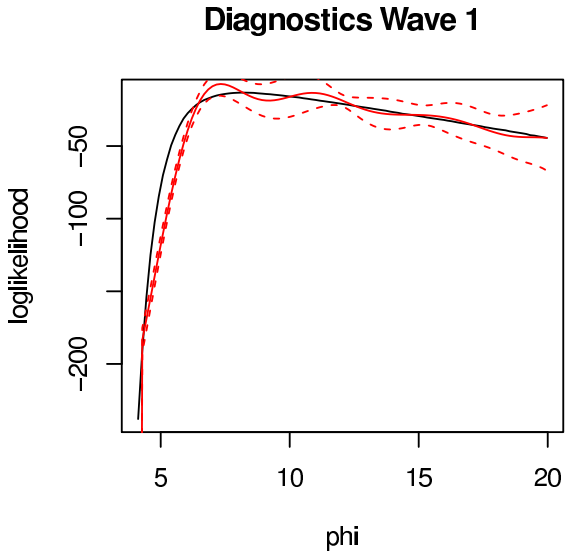


Diagnostics for GP 2 modelling log $l(\theta)$

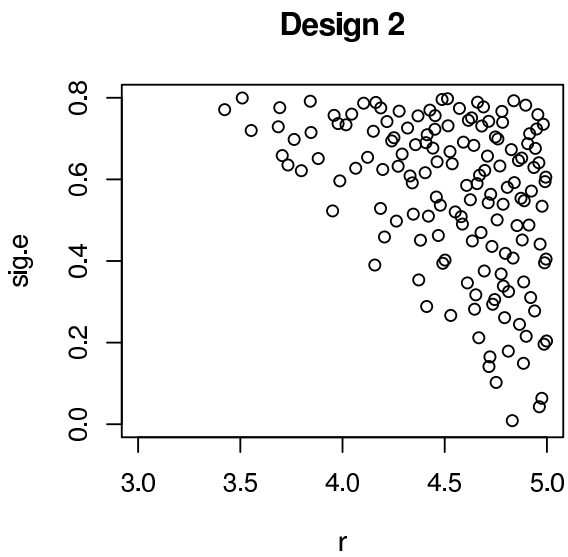
threshold = -21.8

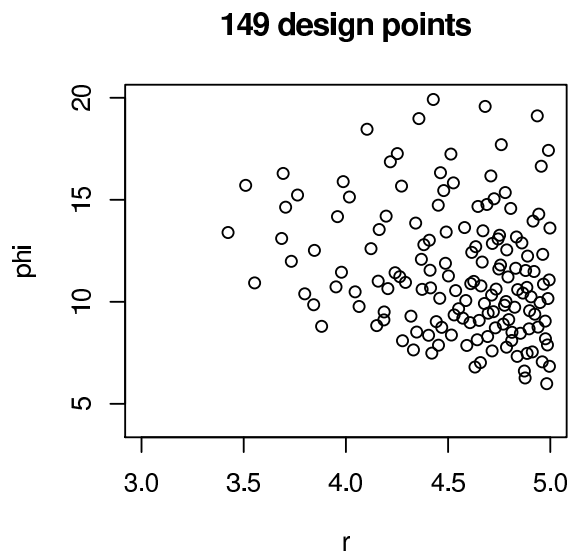


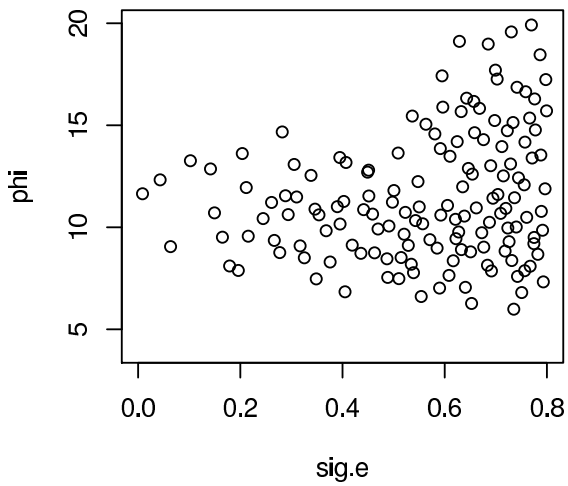




Design 3 - 149 pts - 62% of space implausible

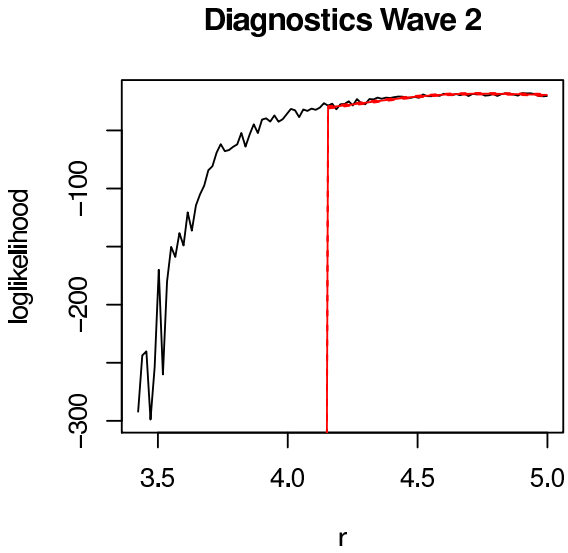


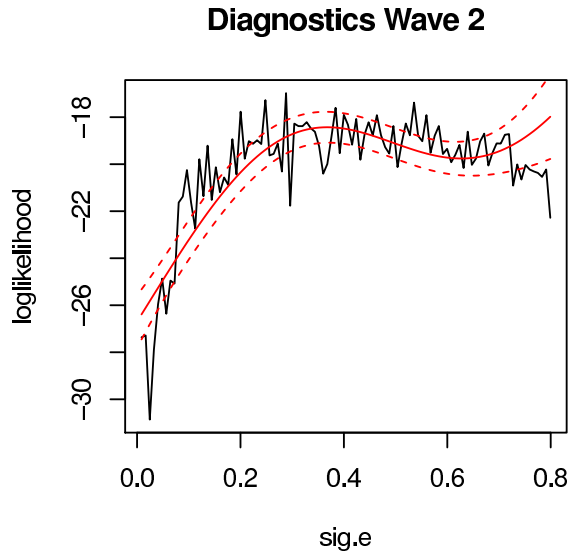


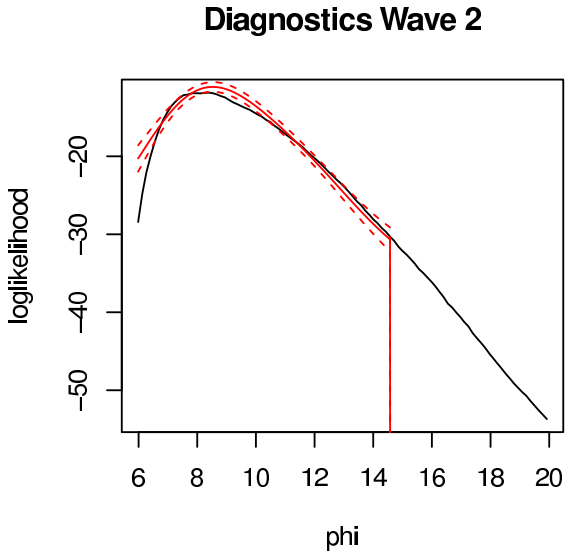


Diagnostics for GP 3 modelling $\log l(\theta)$

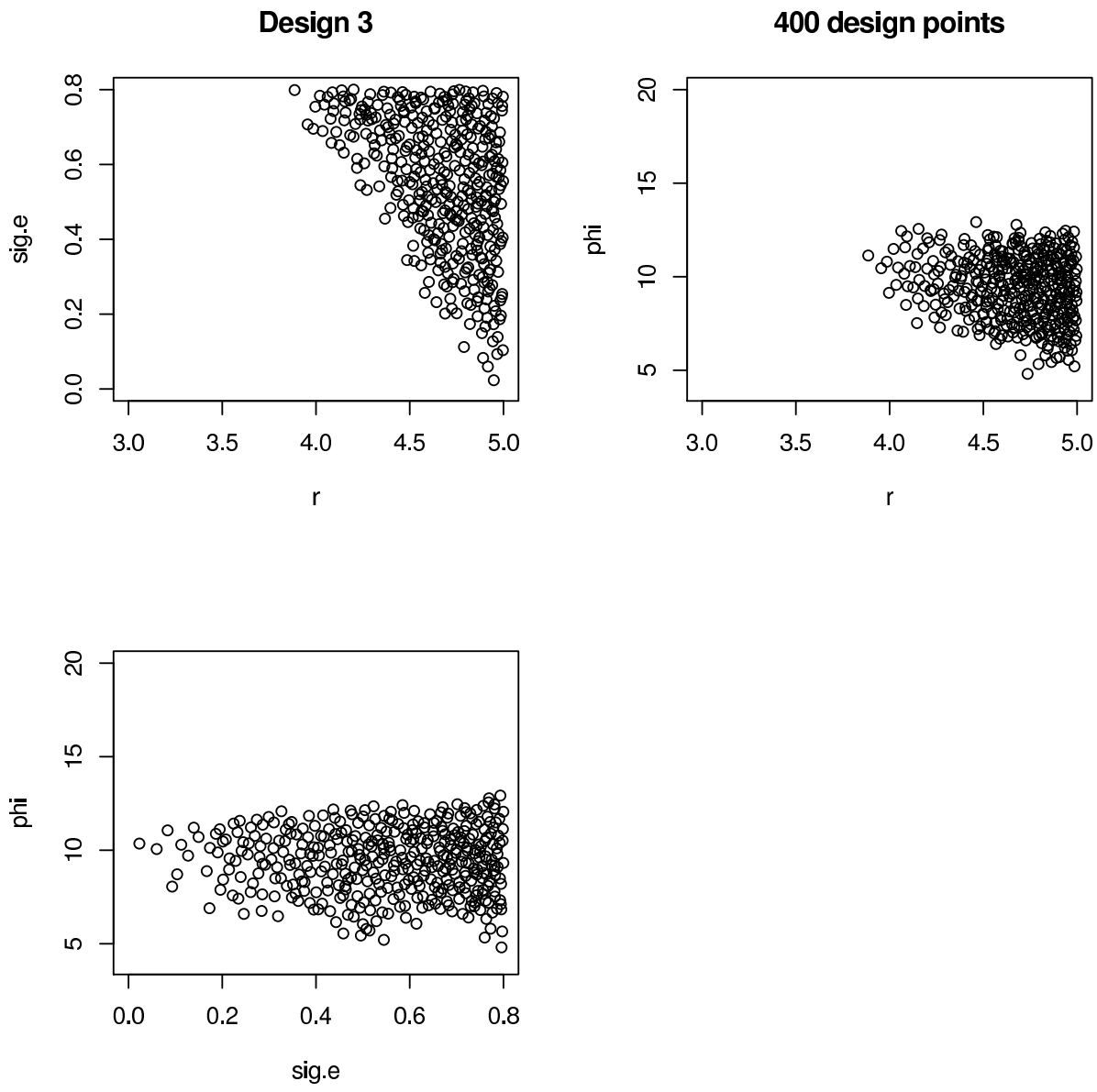
Threshold = -20.7





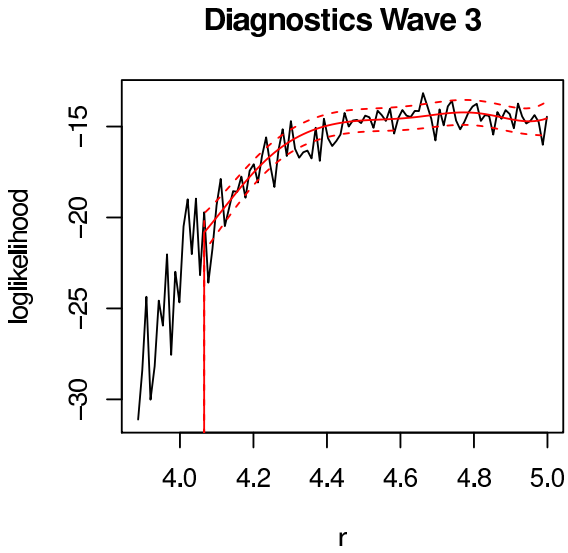


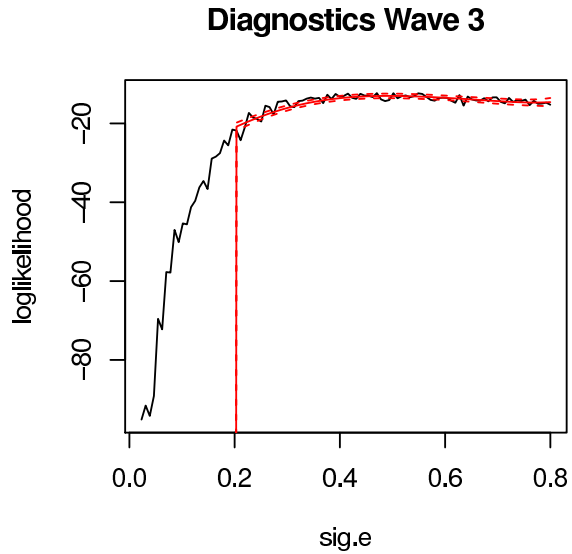
Design 4 - 400 pts - 95% of space implausible

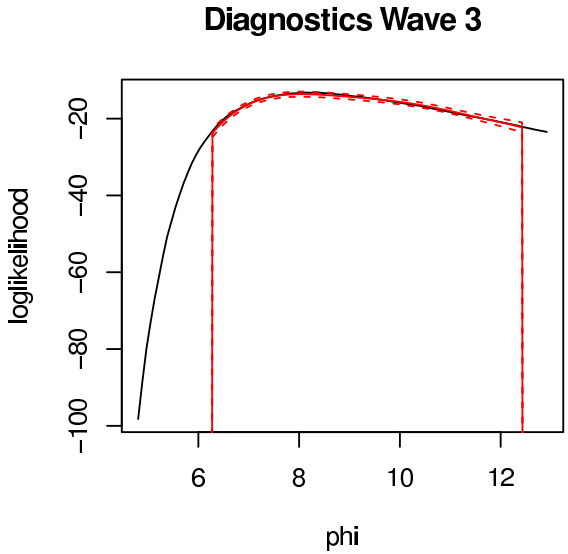


Diagnostics for GP 4 modelling log $l(\theta)$

Threshold = -16.4

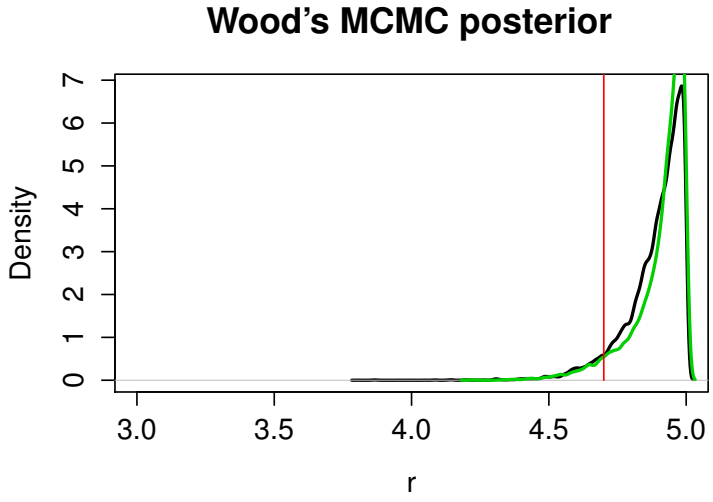


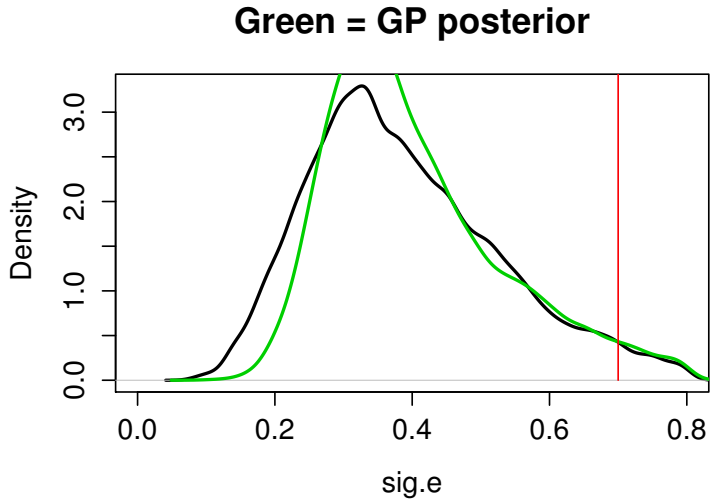


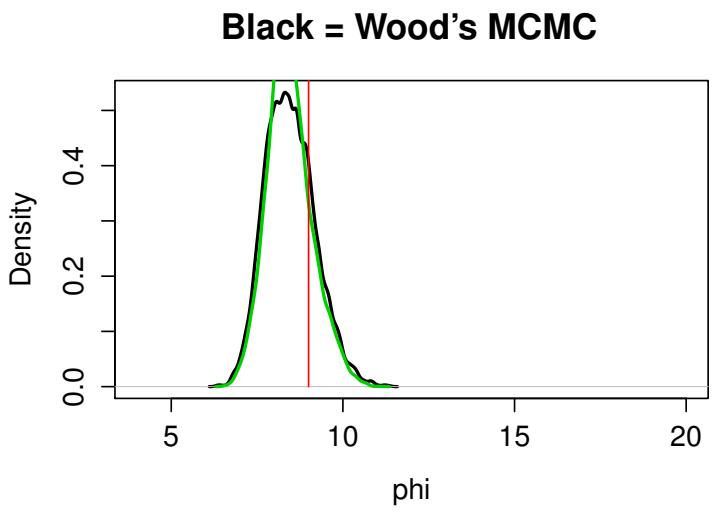


MCMC Results

Comparison with Wood 2010, synthetic likelihood approach







- The Wood MCMC method used $10^5 \times 500$ simulator runs
- The GP code used $(128 + 314 + 149 + 400) = 991 \times 500$ simulator

Acquisition rules

Work with James Hensman

We are classifying space as plausible or not by estimating the probability

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

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$,

Acquisition rules

Work with James Hensman

We are classifying space as plausible or not by estimating the probability

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

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$,
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

Acquisition rules

Work with James Hensman

We are classifying space as plausible or not by estimating the probability

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

where $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$,
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.

To find the MAP estimate we can use standard acquisition rules such as the expected improvement. But we want more than a point estimate.

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.

To find the MAP estimate we can use standard acquisition rules such as the expected improvement. But we want more than a point estimate. First idea: add design points where we are most uncertain

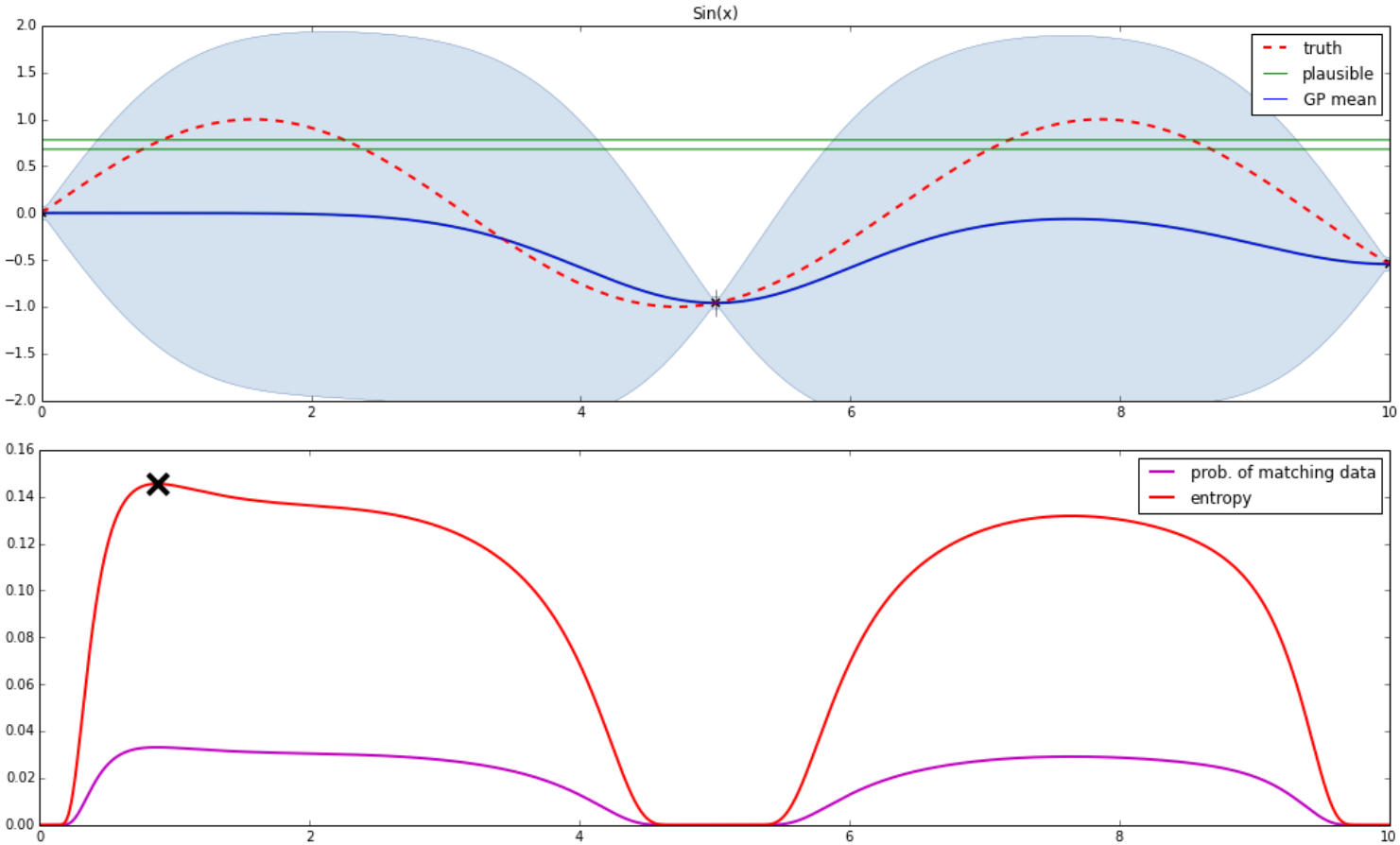
- The entropy of the classification surface is

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

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

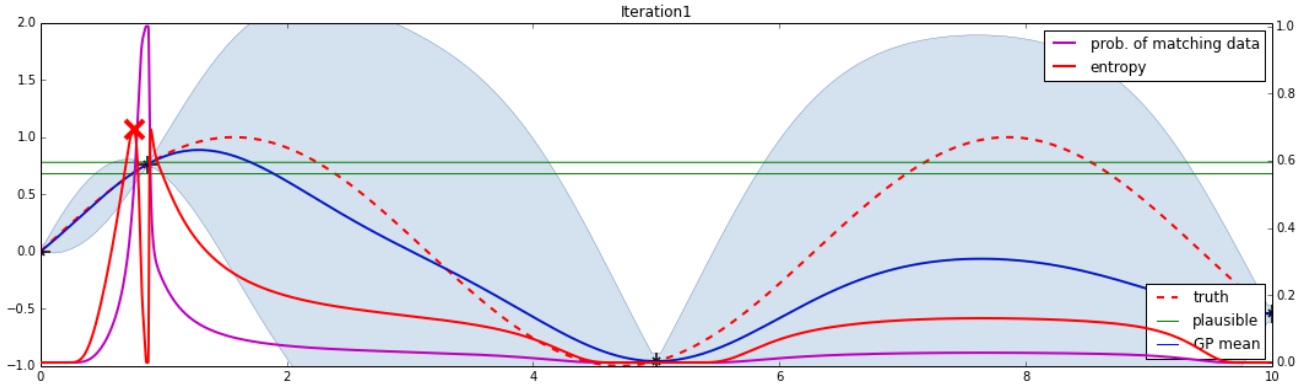
$$\theta_{n+1} = \arg\max E(\theta)$$

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

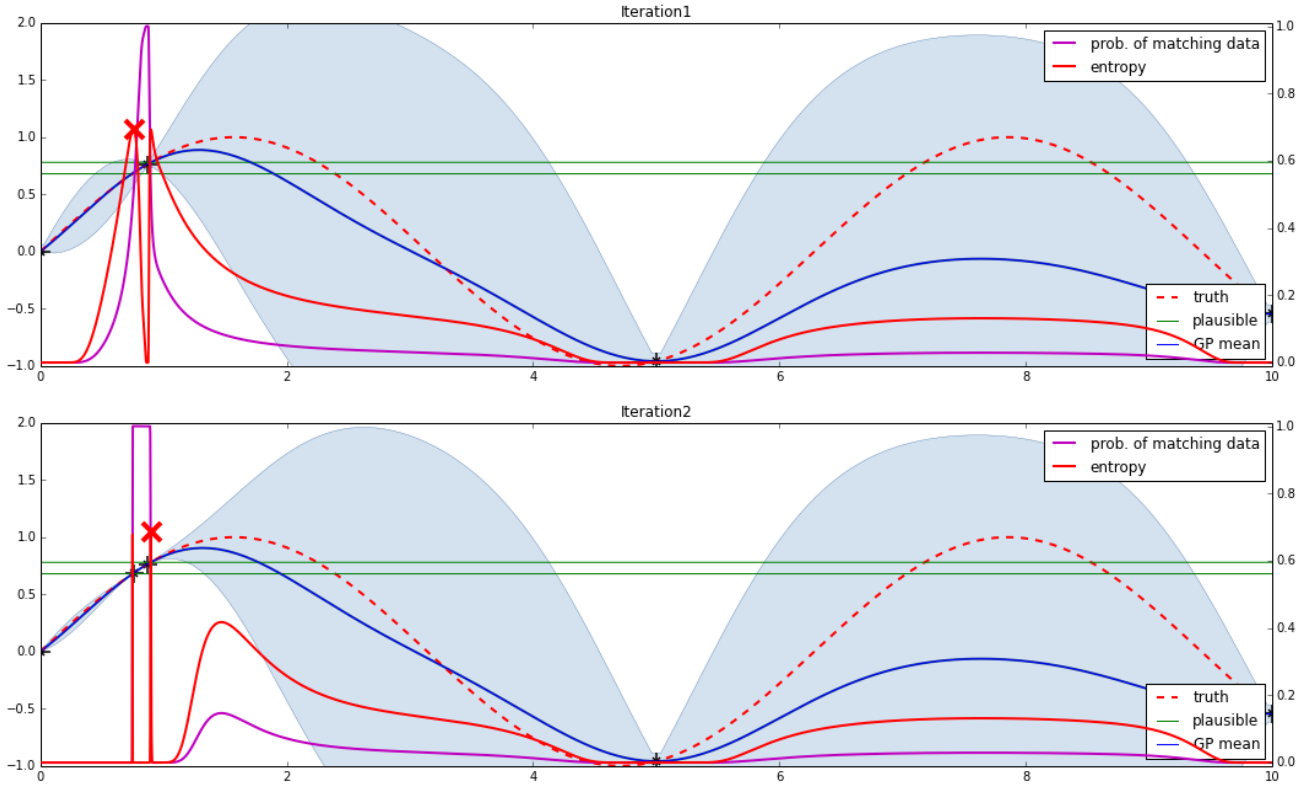


Add a new design point (simulator evaluation) at the point of greatest entropy

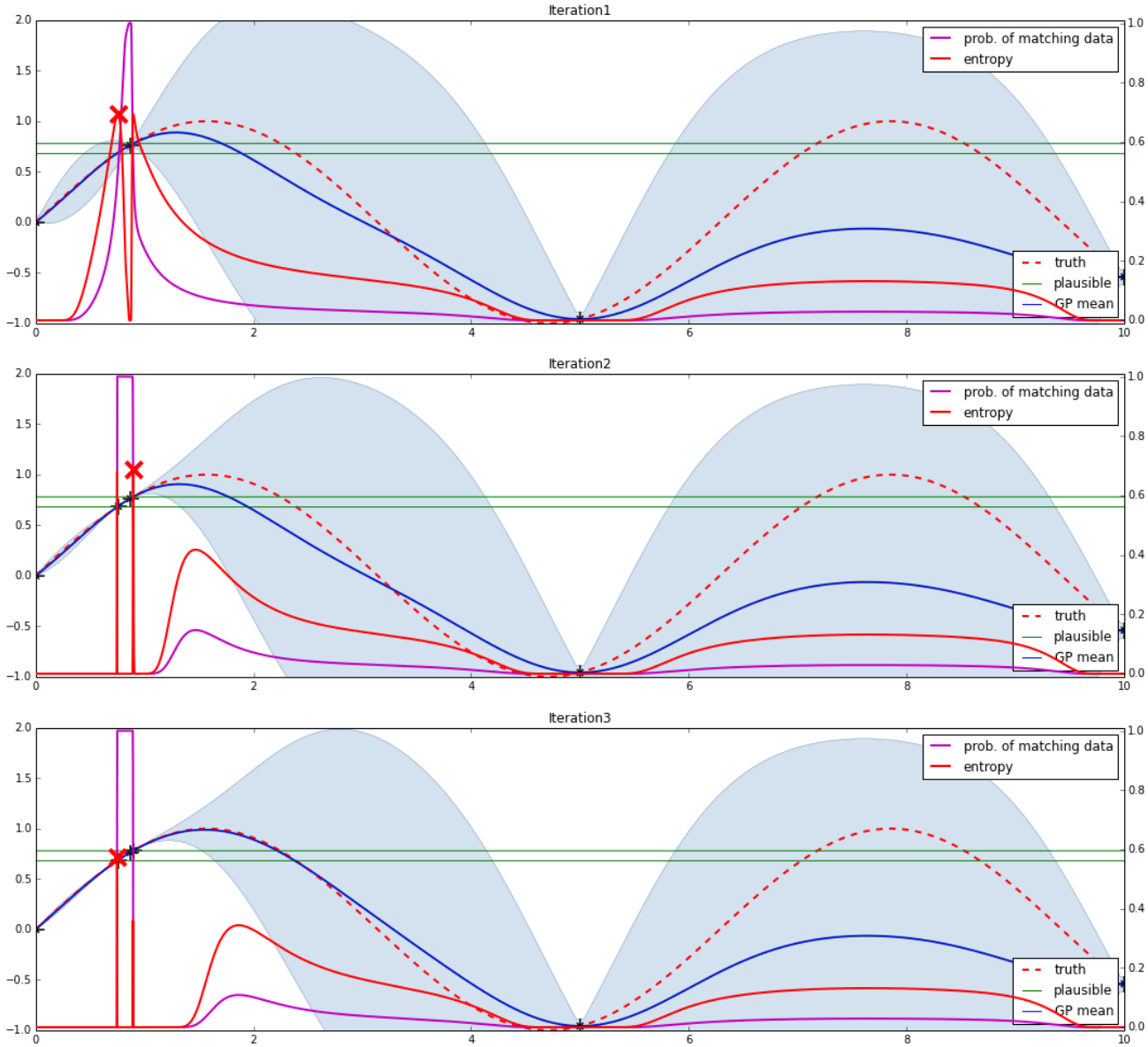
Toy 1d example $f(\theta) = \sin\theta$



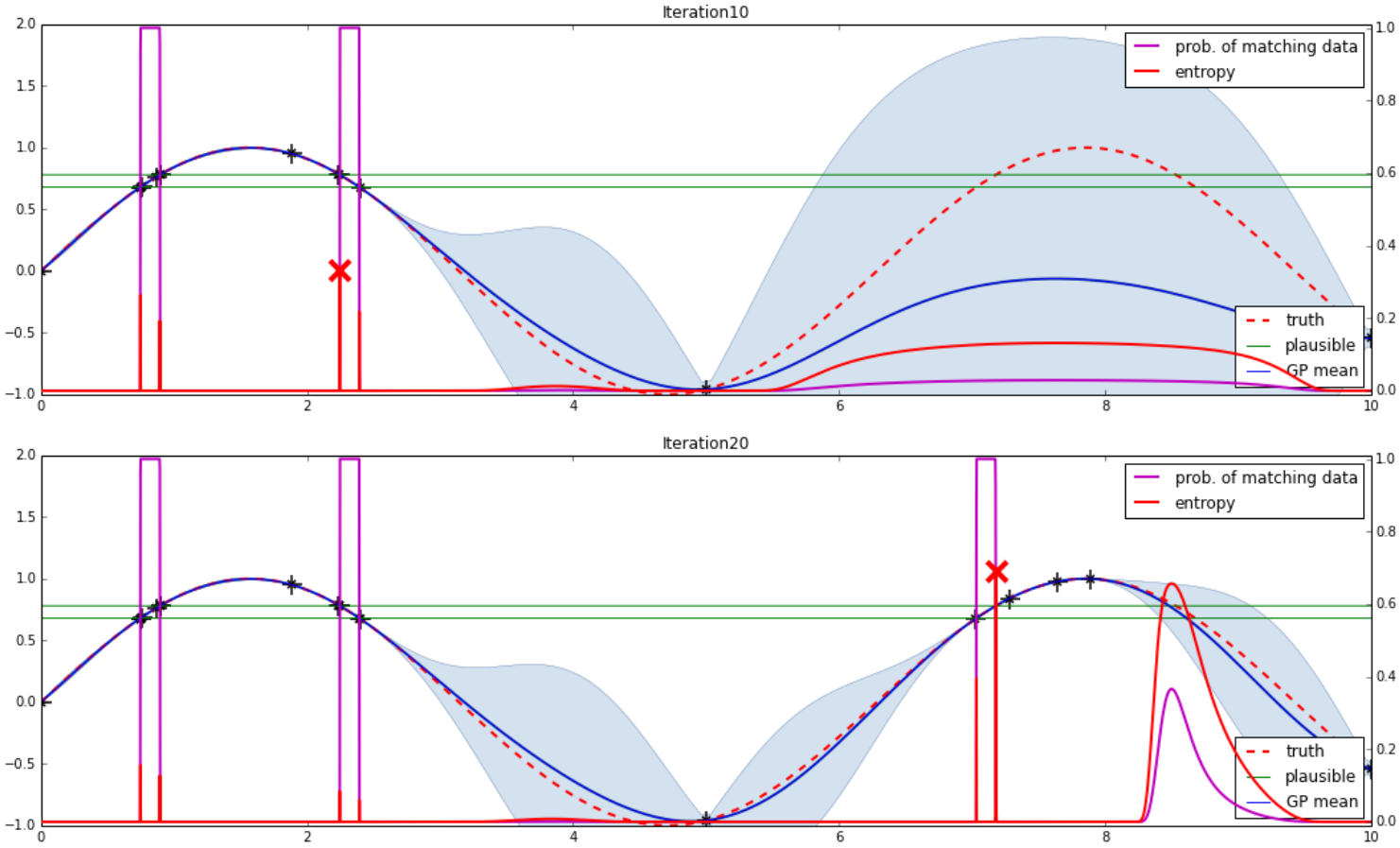
Toy 1d example $f(\theta) = \sin\theta$



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.

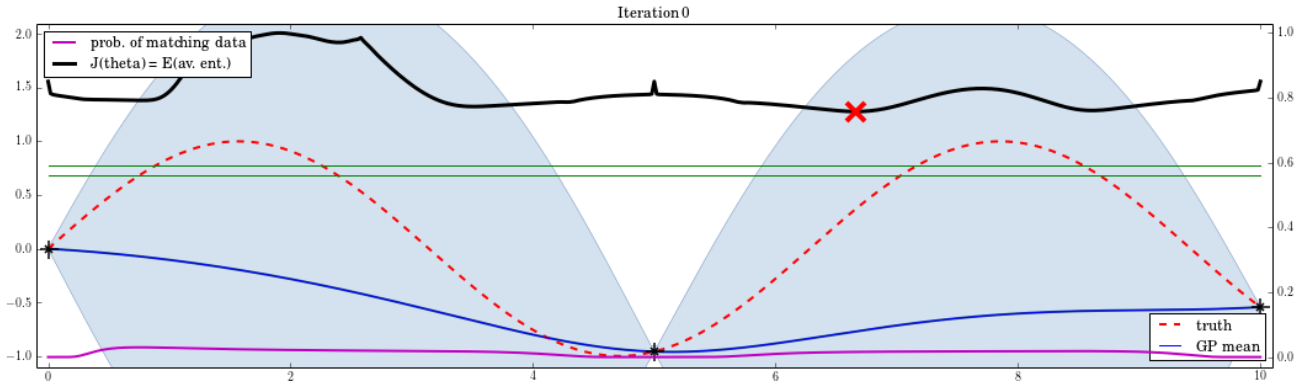
- Choose the next design point, θ_{n+1}, to minimise the expected average entropy

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

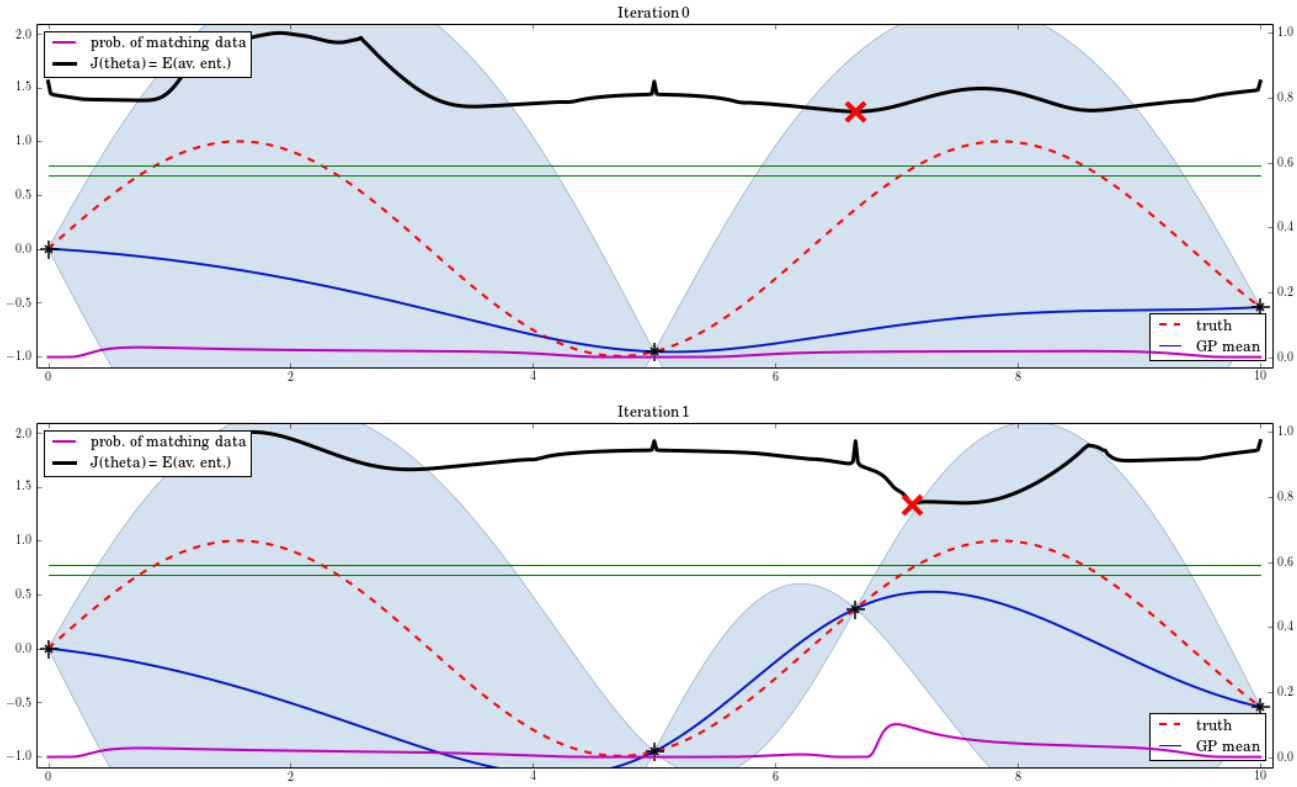
  where

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

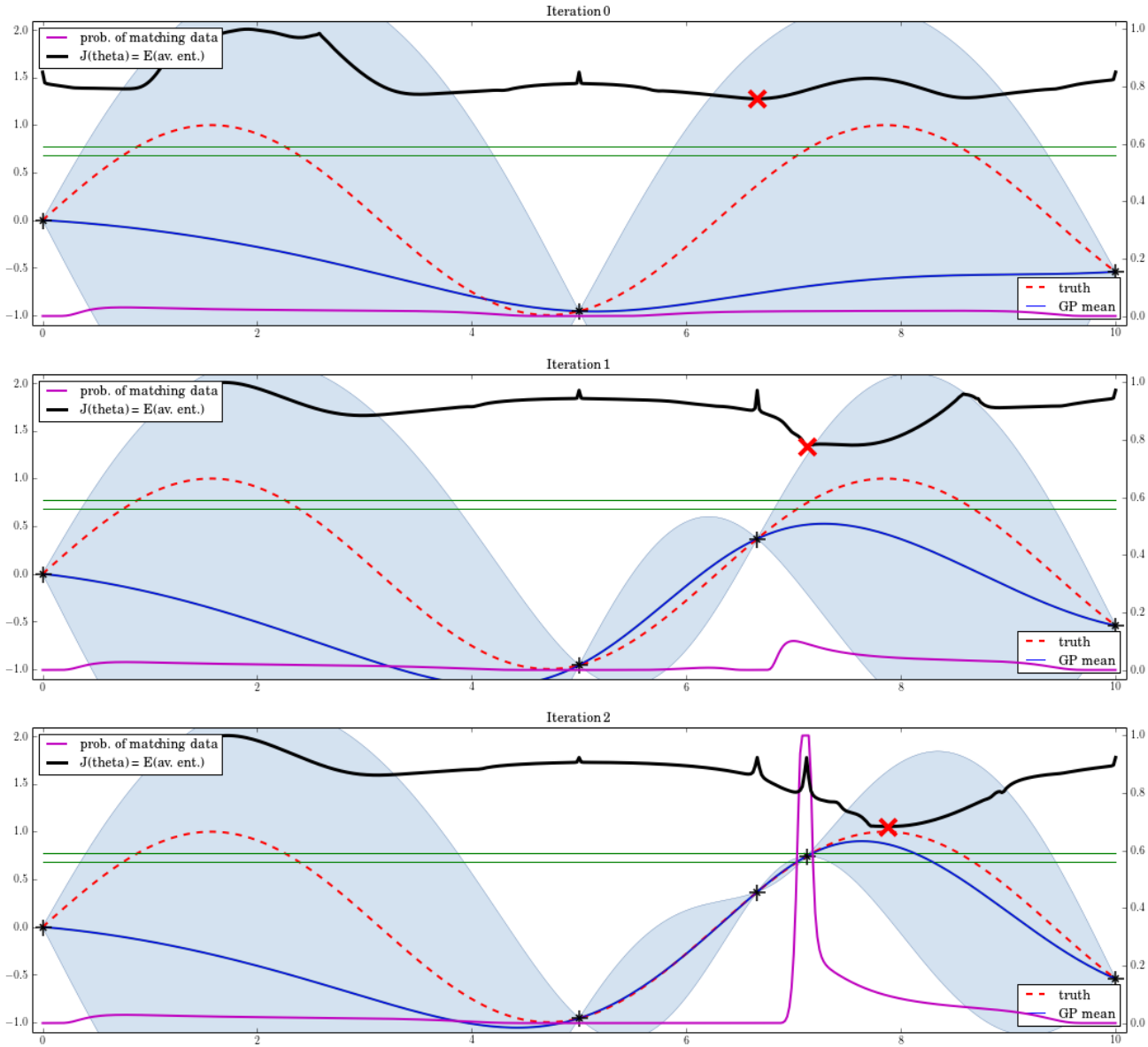
Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



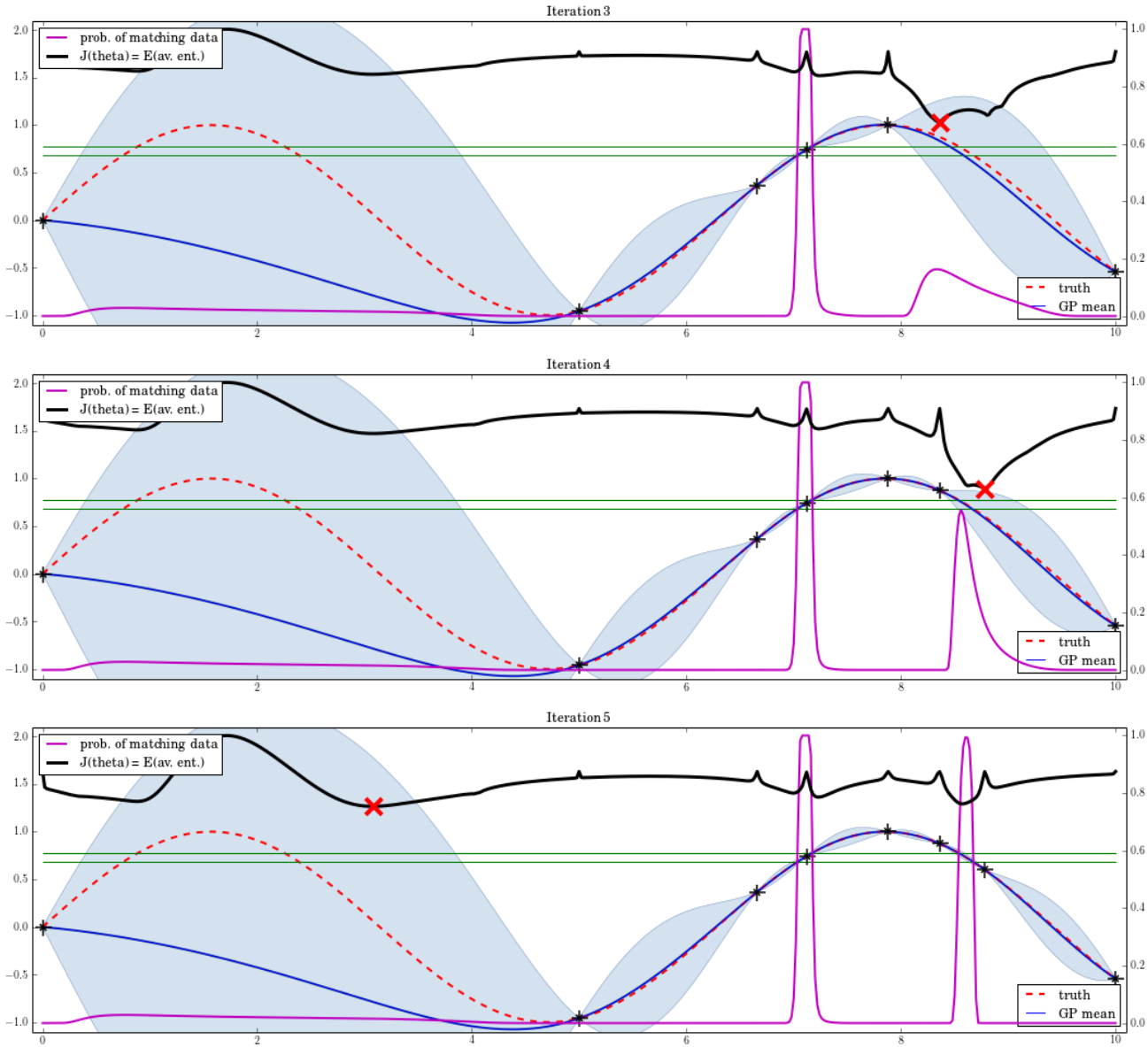
Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



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

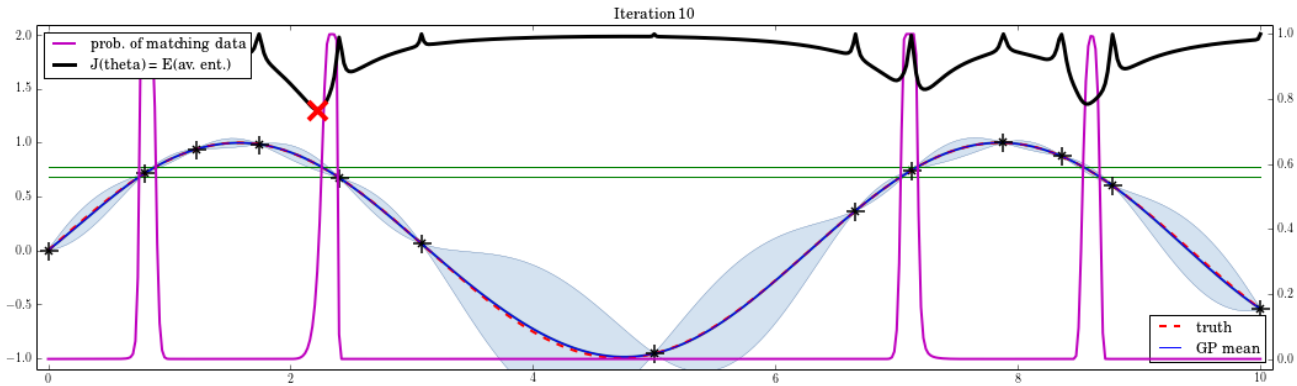


Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



Toy 1d: min expected entropy vs max entropy

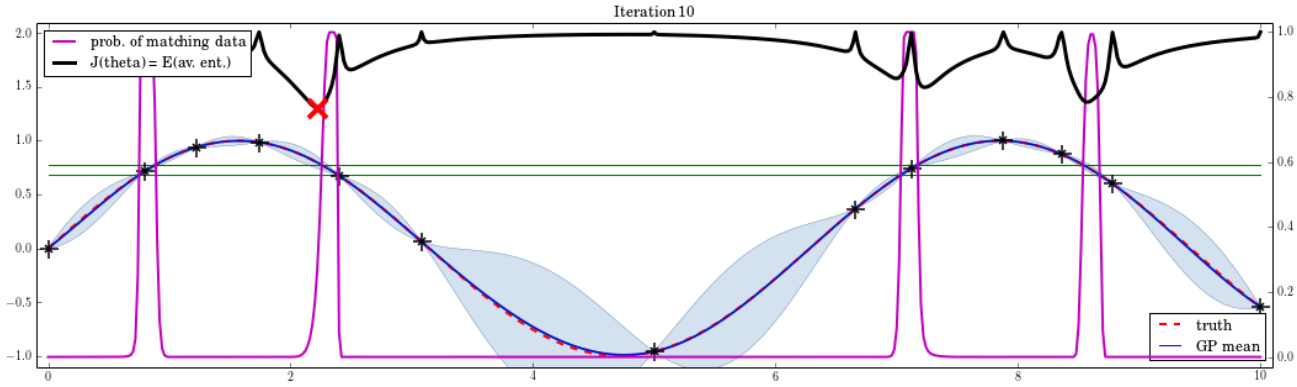
After 10 iterations, choosing the point of maximum entropy



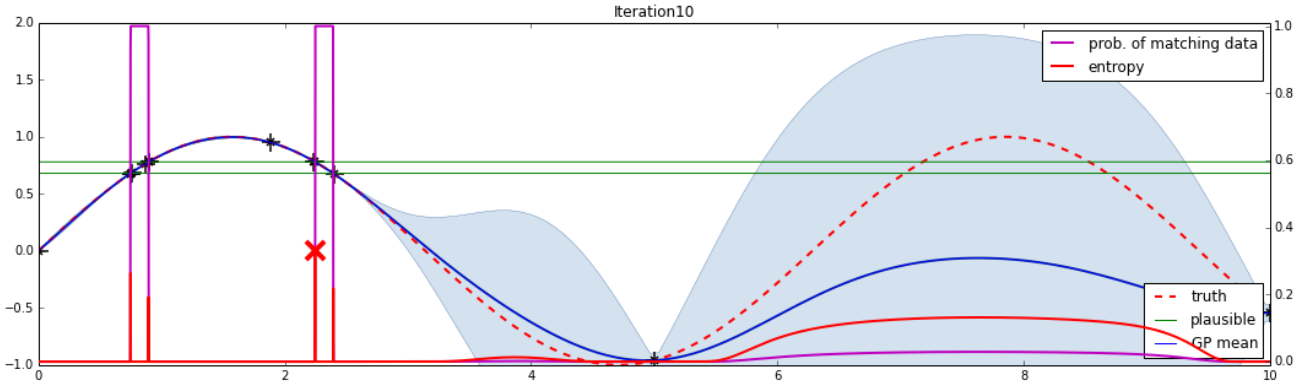
we have found the plausible region to reasonable accuracy.

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

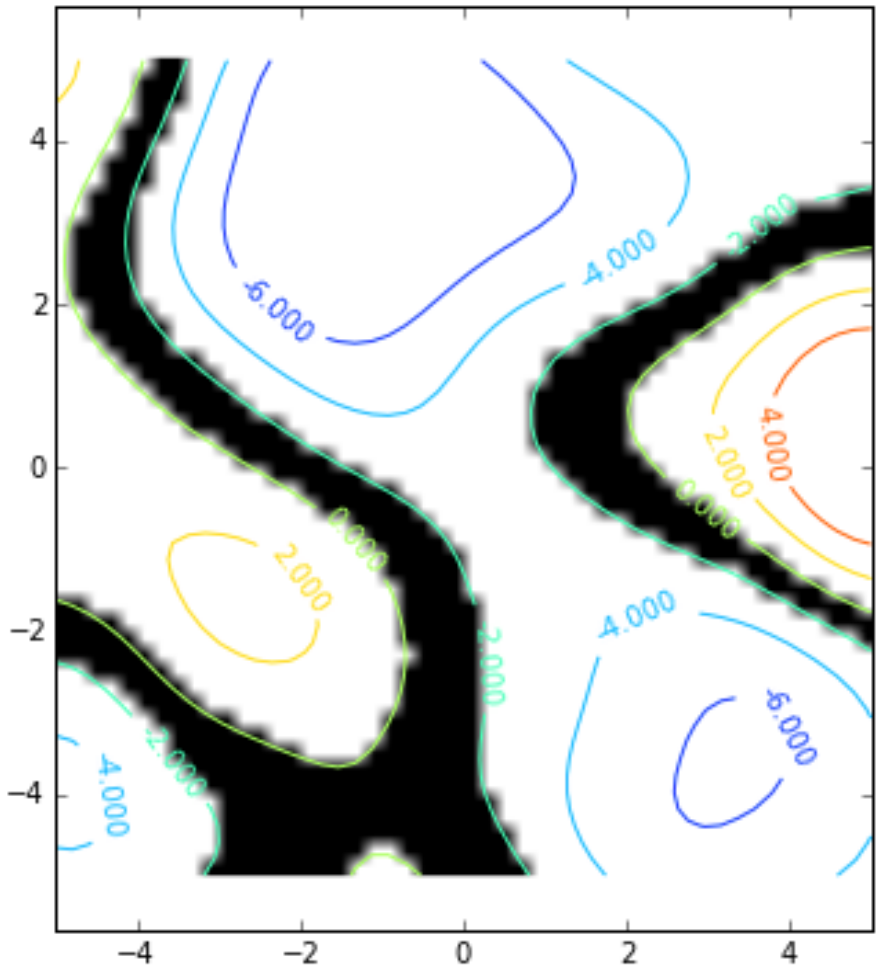
We can use Bayesian optimization to find the optima:

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

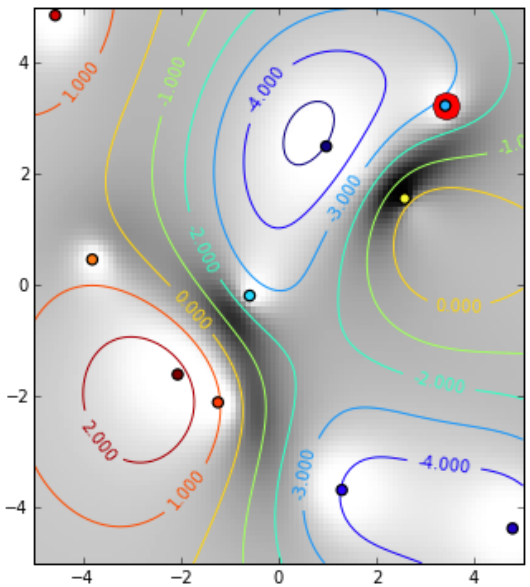
Can we learn the following plausible set?

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



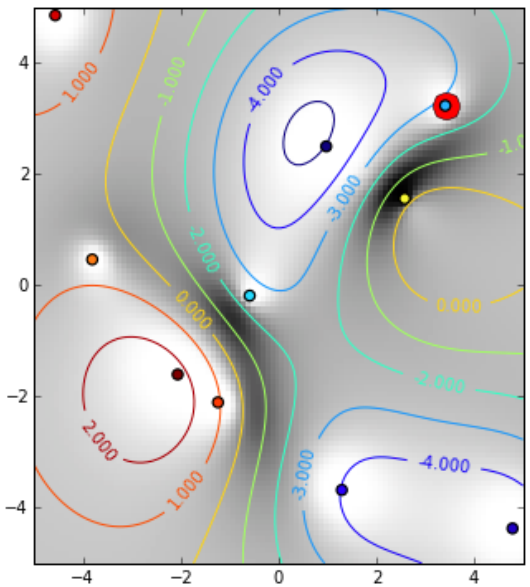
Iteration 10

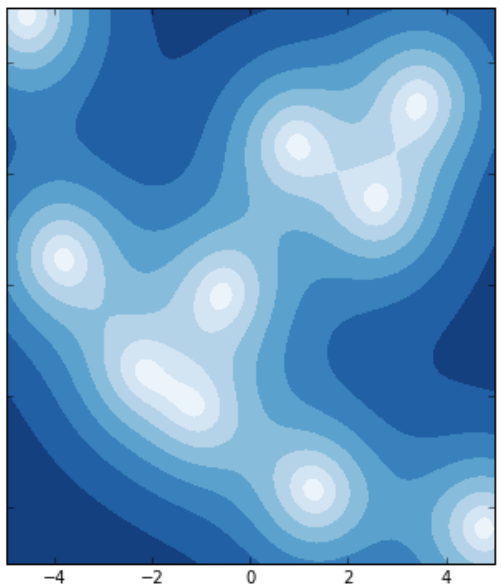
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$



Iteration 10

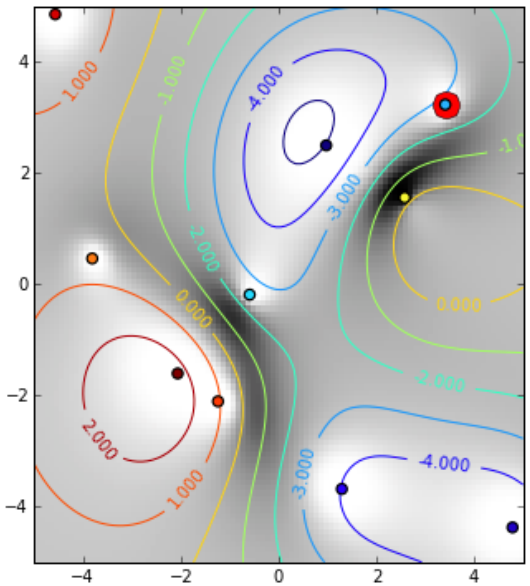
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

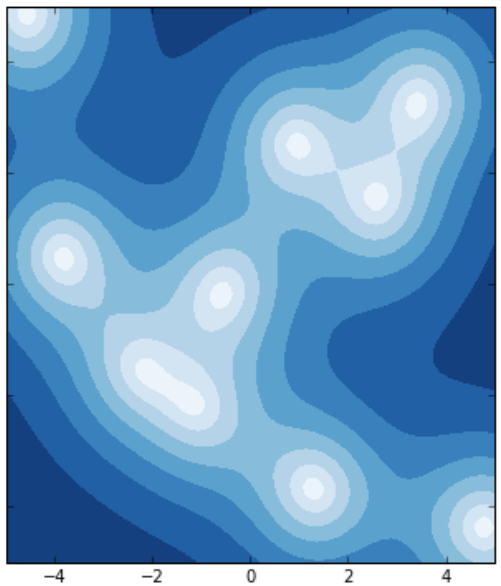


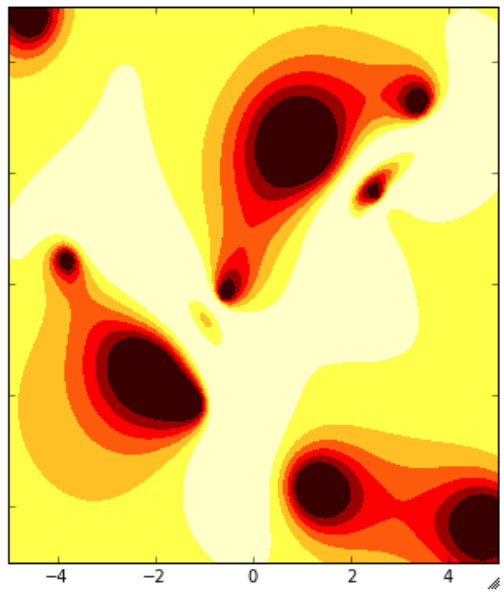


Iteration 10

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

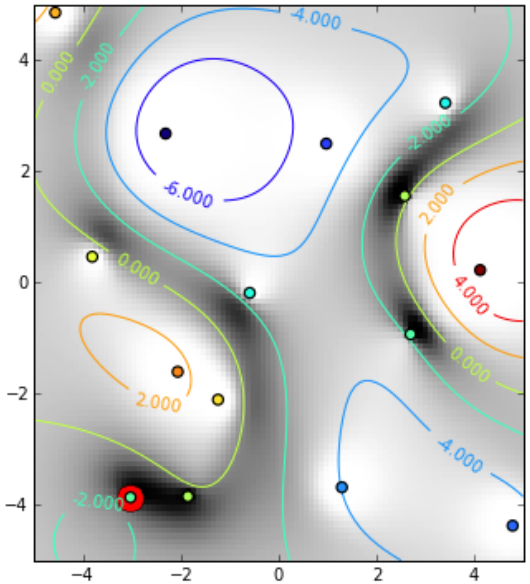


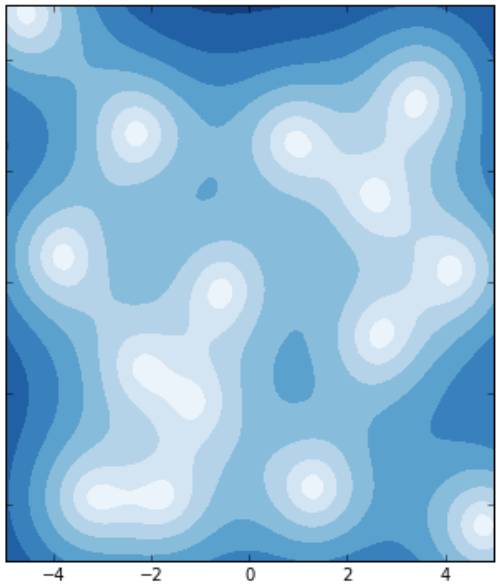


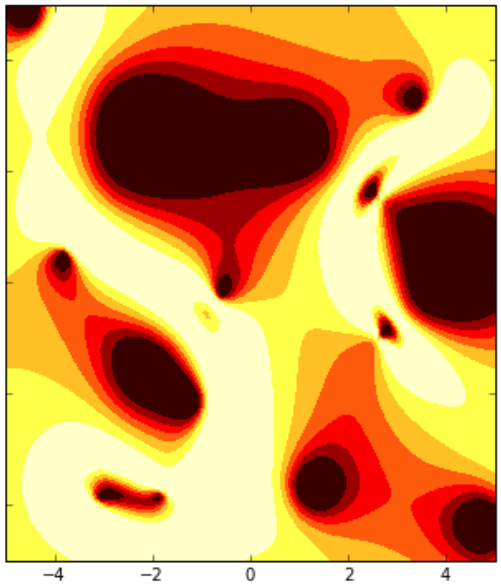


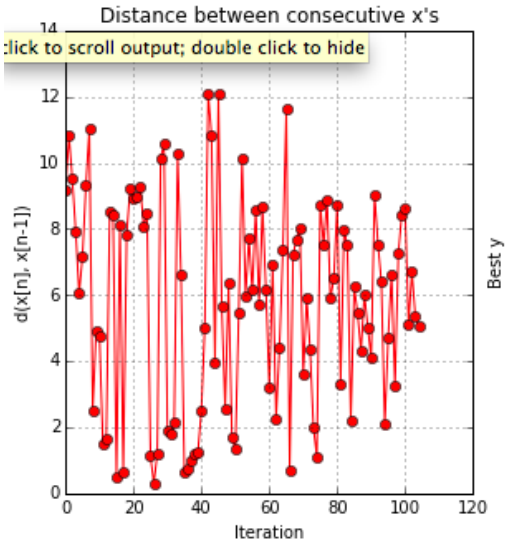
Iteration 15

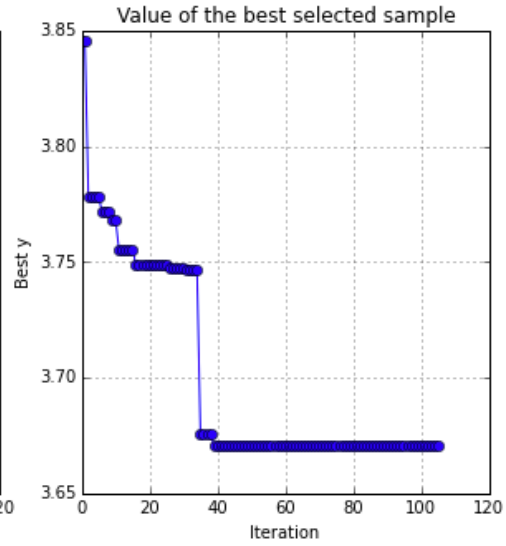
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

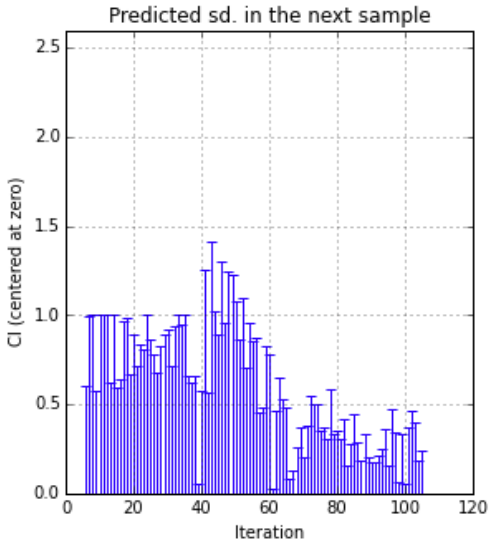




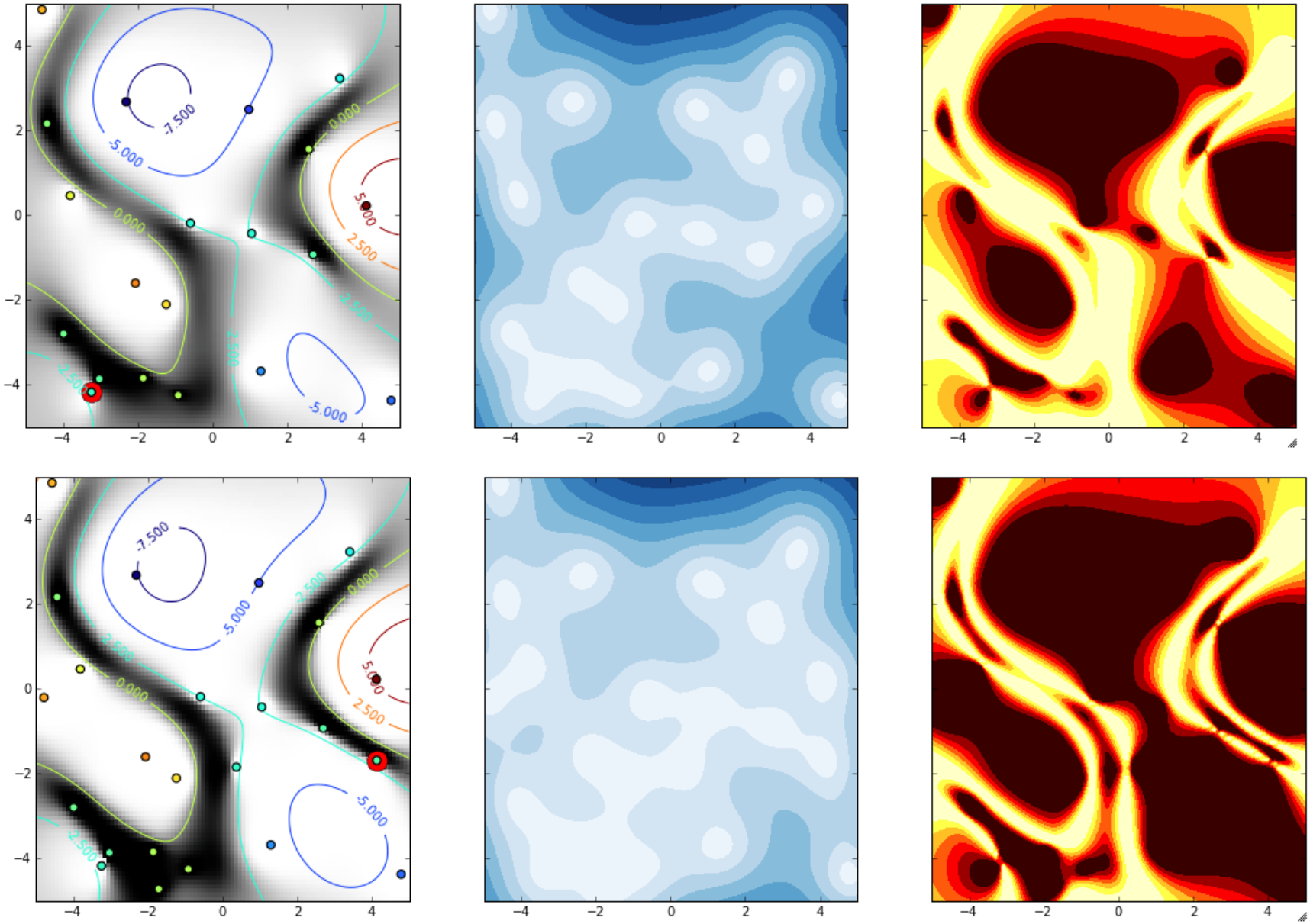








Iterations 20 and 24

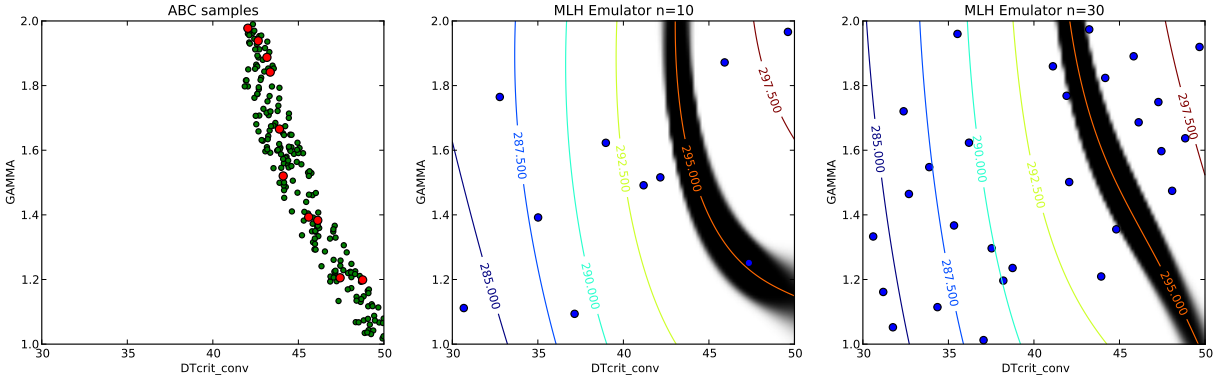


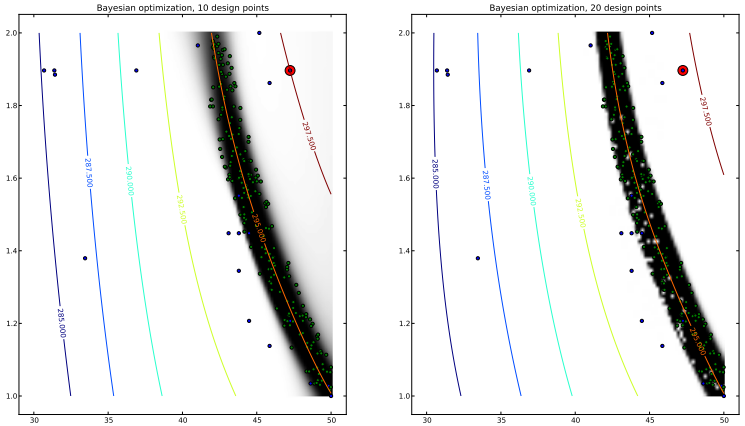
Video

EPm: climate model

Holden *et al.* 2016

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





Inference

- Kennedy and O'Hagan 2001 used the surrogate to calculate the posterior - over-utilizes the surrogate, sacrificing exact sampling.
- Rasmussen 2003 corrected for the use of a surrogate in a HMC scheme using a Metropolis step, which requires simulator evaluations at every stage - under-utilizes the surrogate, sacrificing speed-up
- Sherlock *et al.* 2015 use delayed-acceptance MCMC which also requires one sim run per accepted value.

Inference

- Kennedy and O'Hagan 2001 used the surrogate to calculate the posterior - over-utilizes the surrogate, sacrificing exact sampling.
- Rasmussen 2003 corrected for the use of a surrogate in a HMC scheme using a Metropolis step, which requires simulator evaluations at every stage - under-utilizes the surrogate, sacrificing speed-up
- Sherlock *et al.* 2015 use delayed-acceptance MCMC which also requires one sim run per accepted value.

Conrad *et al.* 2015 use local approximations to produce a MC sampler that asymptotically samples from the exact posterior.

- experimental design combines guidance from MCMC and local space filling heuristics, triggered by random refinement and local error indicators of model quality.
  - 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.
- Allows for rigorous error analysis.

Inference scheme

Is it really necessary to correct for the surrogate in the inference?

George Box 1976

> ~~*All models are wrong but some are useful*~~
>
> *It is inappropriate to be concerned about mice when there are tigers abroad*

We are missing an understanding of what is importantly wrong

- Model error
- sampling errors
- simulator variance
- ABC approximation
- summaries

Problems

- Error analysis: we don't want to spend too long achieving accuracy we don't need. Given the model error, MC error, stochastic variance of the simulator, how much effort should we spend on refining the surrogate?
- Design/acquisition: need a batch acquisition rule that accounts for likelihood-estimate errors and surrogate errors.
- Simulator discrepancy
- Combining this approach with methods to find good summaries of the output $S(D)$.

Conclusions

- For complex models, surrogate-modelling approaches are often necessary
- Target of approximation: discrepancy vs likelihood vs simulator output
- 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
- Still much to do...

Conclusions

- For complex models, surrogate-modelling approaches are often necessary
- Target of approximation: discrepancy vs likelihood vs simulator output
- 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
- Still much to do...

Thank you for listening!