Surrogate modelling and ABC

Richard Wilkinson

University of Sheffield

Motivation

Expensive stochastic simulators exist

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.
- 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.

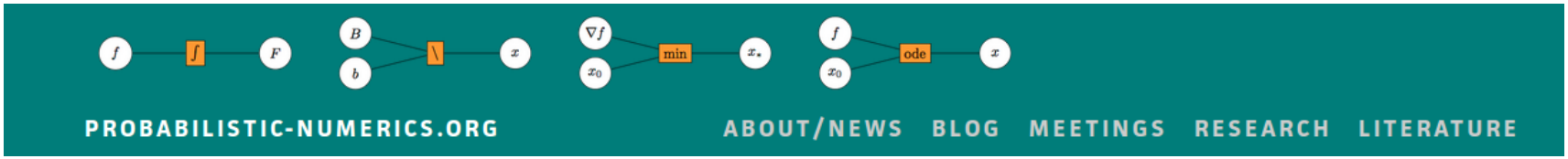
Efficient algorithms can take us only so far...

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

Outline

- Probabilistic numerics
- Surrogate ABC
  - Target of approximation
  - Aim of inference
  - Surrogate model
  - Acquisition rule

Uncertainty quantification and Probabilistic numerics



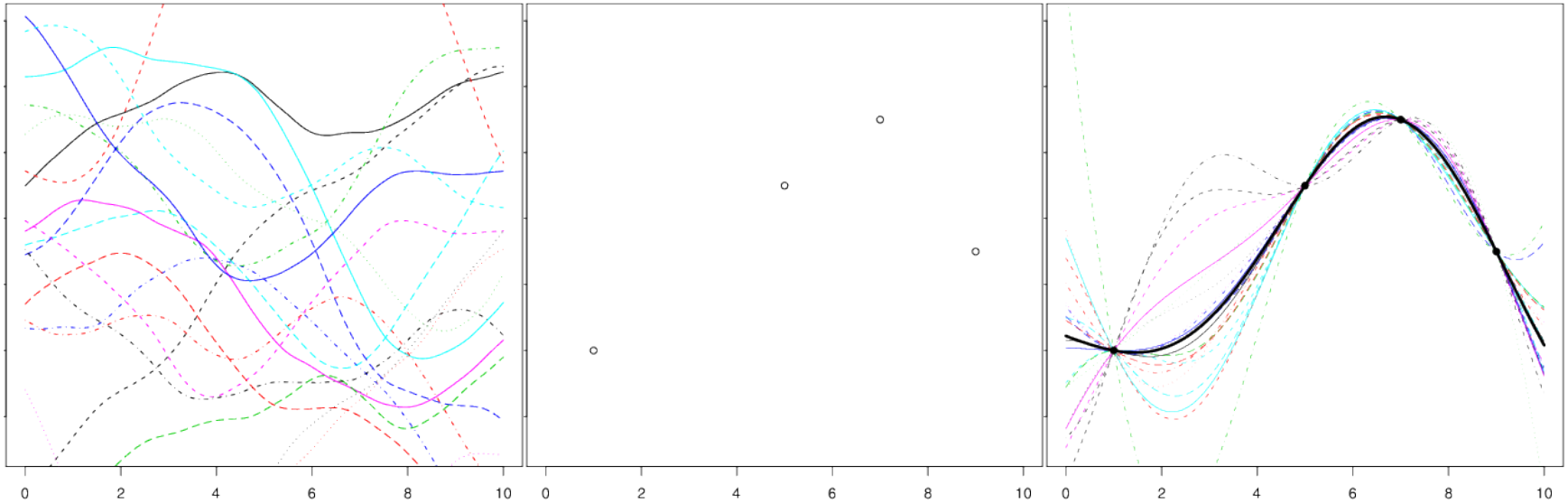
Numerical algorithms, e.g. integration, solving O/PDEs, optimization, estimate some unknown quantity on the basis of function evaluations, ie, they are inference problems.

> *"PN focusses on the computations used to solve a particular problem: what is the uncertainty added by performing the computation approximately?"*

> *"PN offers the attractive potential of performing management of systems of probabilistic numerical algorithms. That is, PN could be used to select which part of a numerical pipeline to refine, that is, to decide when to stop a numerical algorithm achieving accuracy you don't need."*

"Monte Carlo is fundamentally unsound" O'Hagan 1987

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.
  Model f as a GP and add new design points/function queries using some acquisition rule such as expected improvement.

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 simulator, approximate it by a cheaper surrogate model (if in doubt...)

Kennedy and O'Hagan 2001 consider using emulators for 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

Others have done uncertainty analysis, sensitivity analysis, design, error estimation etc.

MCMC

Rasmussen 2003 introduces the idea of using GPs in Hamiltonian Monte Carlo.

- artifical dynamics based on the derivative of

$$E_{pot}(\theta) \propto -\log p(\theta|D)$$

- model $E_{pot}(\theta)$ as a GP. Because the derivative of a GP is also a GP, we are able to generate cheap candidate values of θ
- correct proposals with Metropolis acceptance step

Fielding, Nott, Liong 2011, extend this approach to the case of multi-modal posteriors using tempering.

ABC

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.

ABC

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*

Some surrogate-model ABC papers

- Henderson et al 2009
- Meeds and Welling 2014
- Wilkinson 2014
- Jabot 2014
- **Gutmann and Corander 2015**
- +Others (apols)

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

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

Target of approximation

What should we approximate with the surrogate model?

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

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

Target of approximation

What should we approximate with the surrogate model?

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

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

  - 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
  - interpretable as a statistical model, i.e., $D = X + e$ where $e \sim K_\epsilon(\cdot)$
  - 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

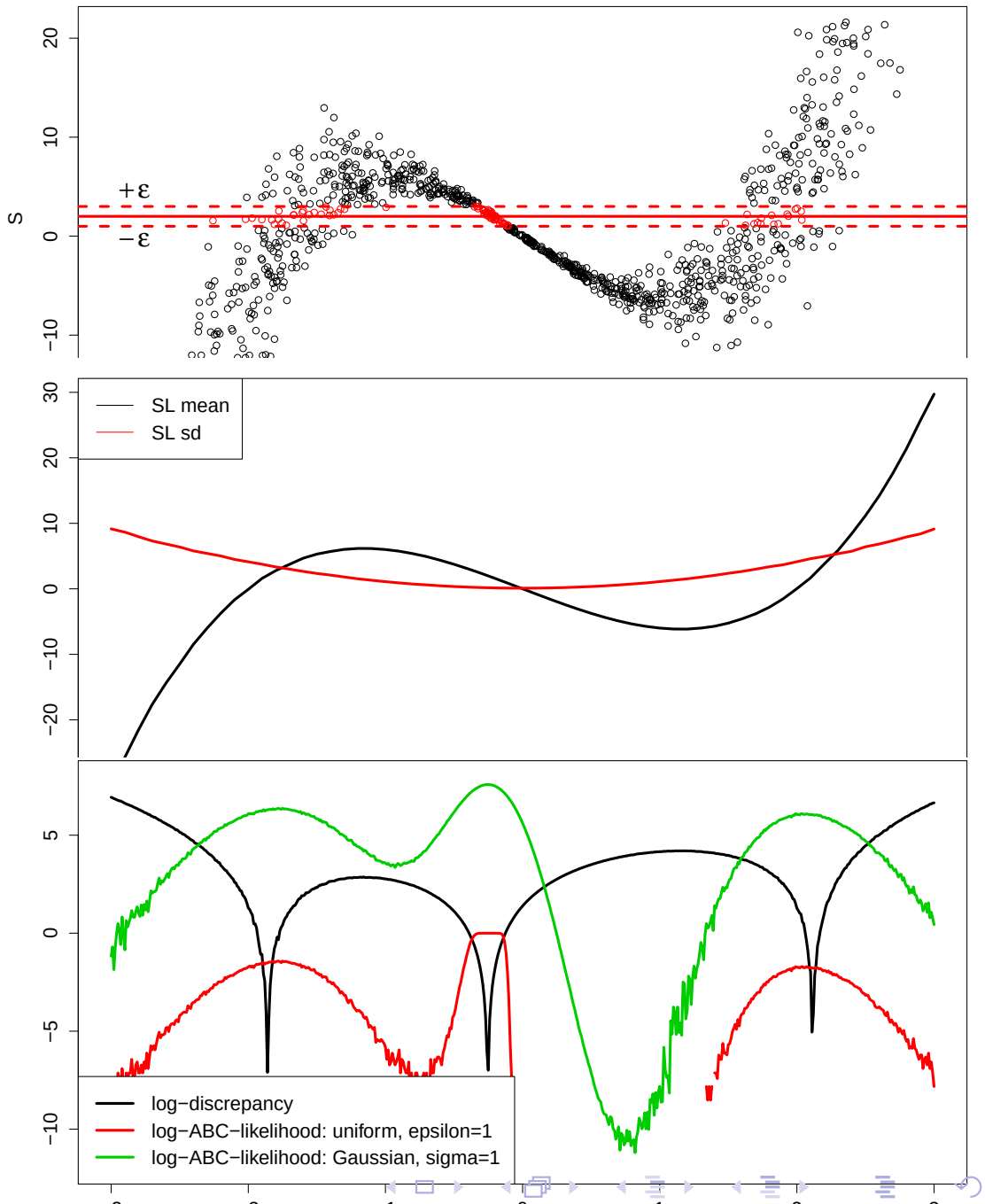
$$J(\theta) = \mathbb{E}\rho(S(D), S(X))$$

  - 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:



The importance of being Bayesian?

Bayesian statistics is sometimes known as inverse probability:

- Unknowns (θ) are given distributions; link to observables using forwards models $X = f(\theta)$, and we use Bayes theorem (inverse probability) to find $\theta|X$.

  E.g., for linear models

  $$y = ax + b + N(0, \sigma^2)$$

  to learn x given y_{obs}, we give a prior to x and then infer $\pi(x|y)$
- Surrogate model approaches aim to preserve this interpretability.

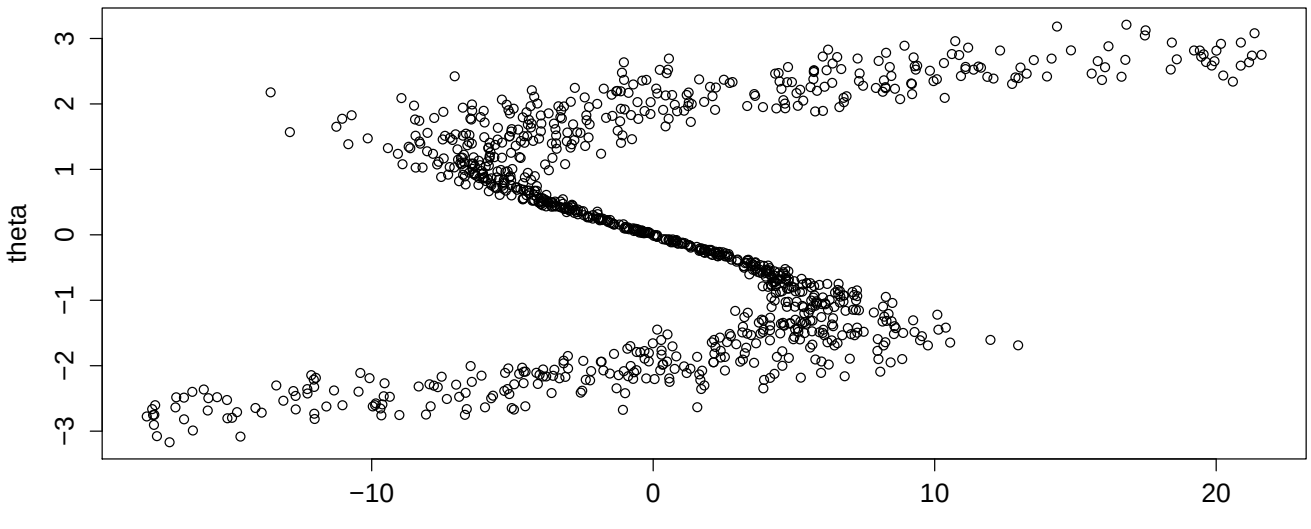
The importance of being Bayesian?

Inverse modelling, as opposed to inverse probability, directly models from observable to unknown:

$$x = a'y + b' + N(0, \sigma'^2)$$

and predict x at y_{obs} as $a'y_{obs} + b'$

- Beaumont *et al.* 2003, Blum and Francois 2010, ... Marin *et al.* 2016? build a model from S to θ

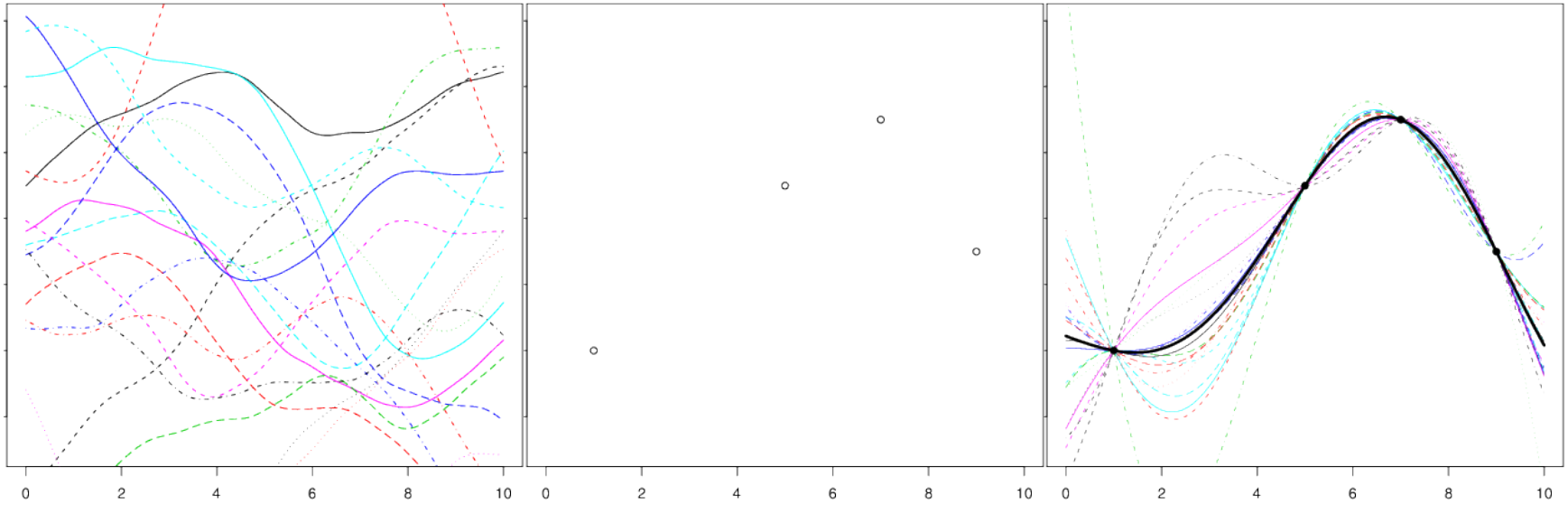


How much do we care about the B in ABC? Is there anything wrong with an inverse modelling approach?

- interpretability of uncertainties?

Choice of surrogate model

- If in doubt...



- Conrad *et al.* 2015, Jabot 2014 use local-linear regression. Find some advantages in terms of tractable error analysis and computational tractability for no degradation in performance. Can't be used in an active learning.
- Sherlock *et al.* 2014 use k-nearest neighbour

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.

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 $\log \pi(D|\theta)$

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

- 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

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.

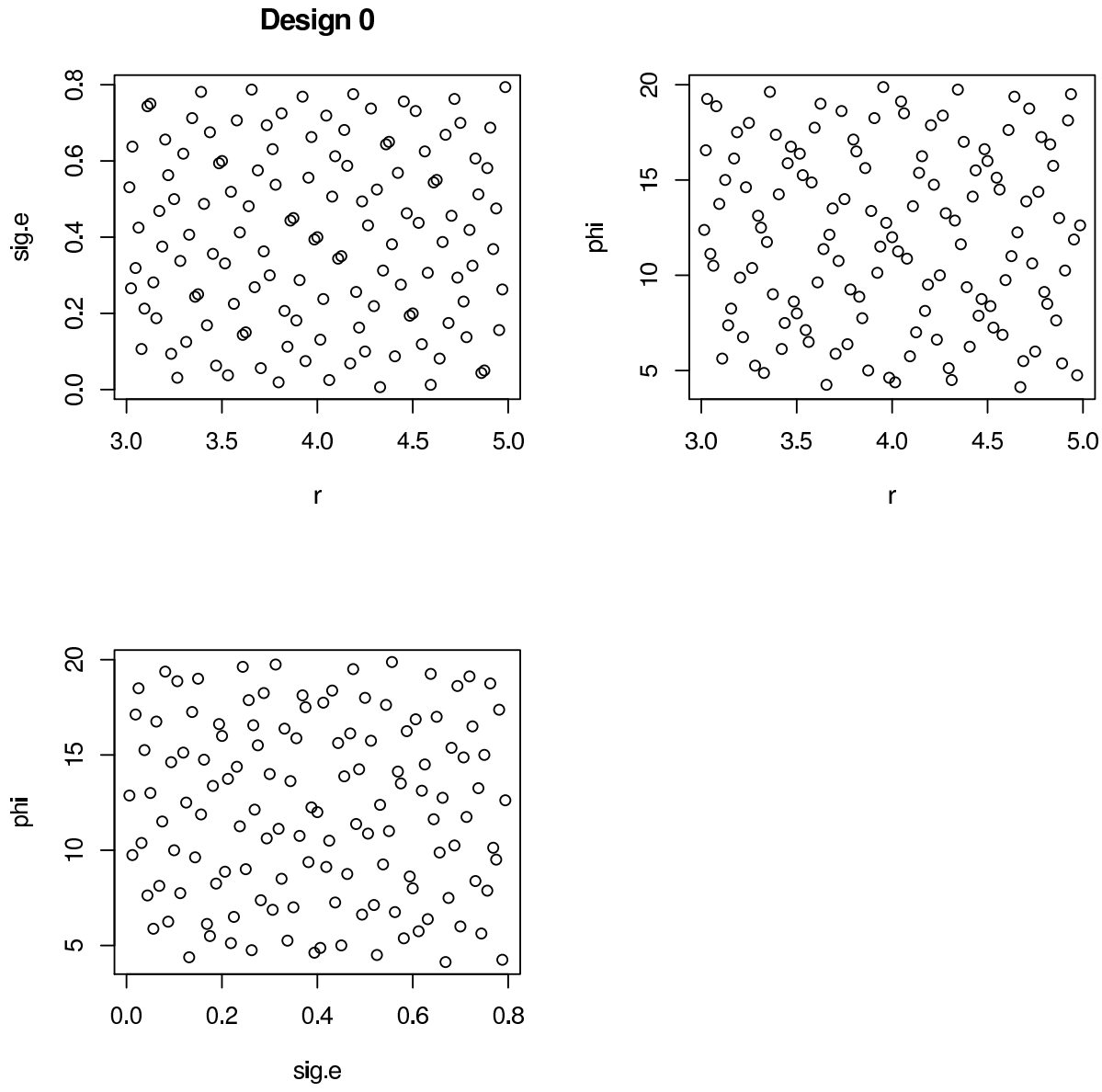
$$N_{t+1} = rN_t e^{-N_t + e_r}$$

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

- Assume we observe counts y_t where

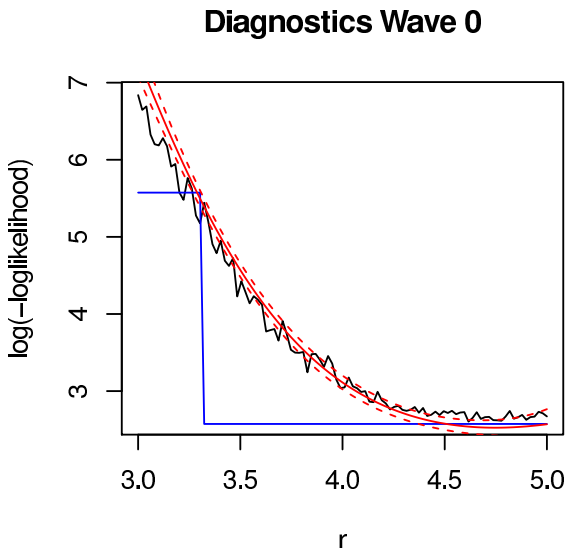
$$y_t \sim Po(\phi N_t)$$

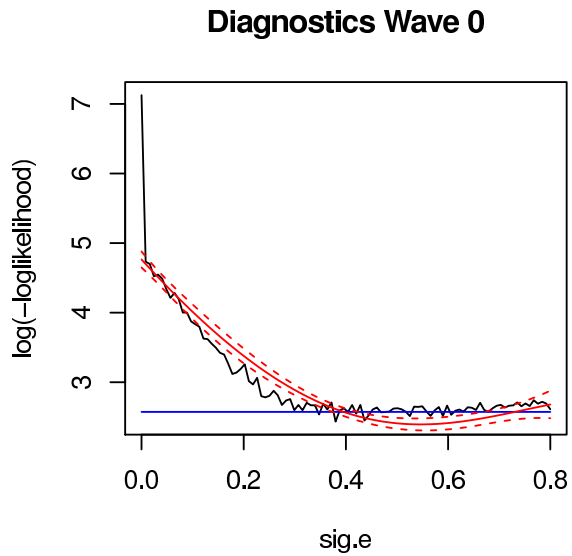
Results - Design 1 - 128 pts

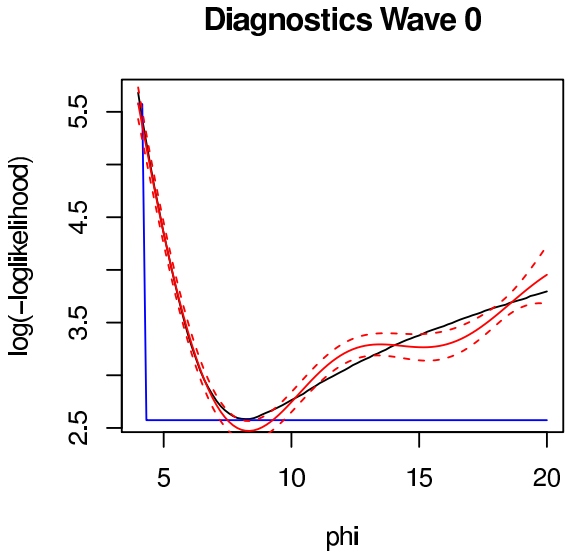


Diagnostics for GP 1 modelling $\log(-\log l(\theta))$

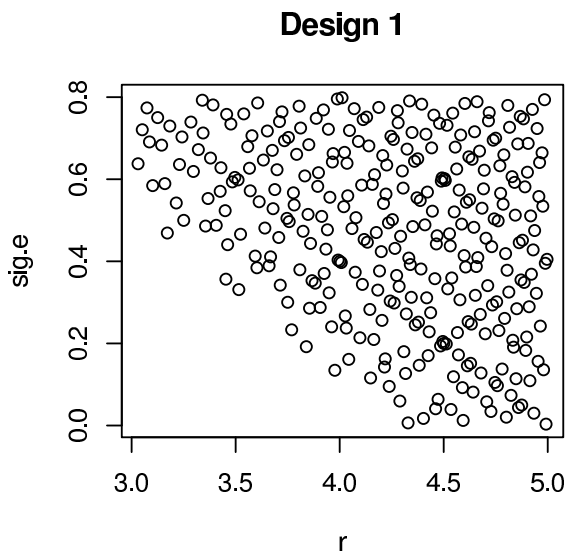
Threshold = 5.6

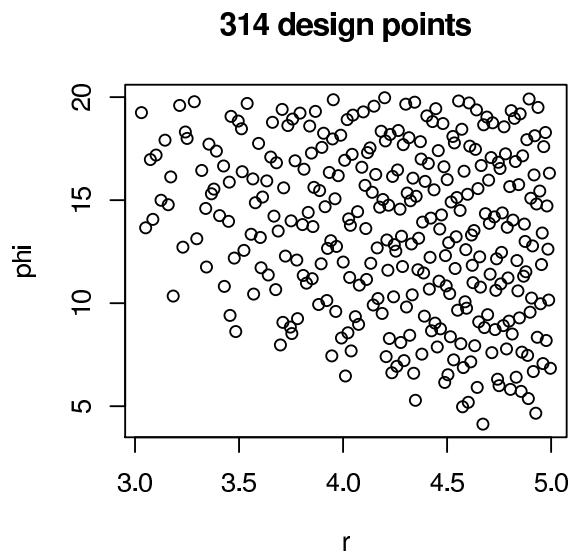


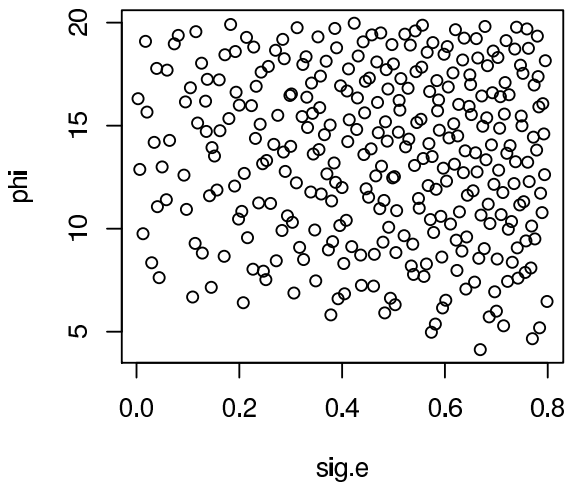




Results - Design 2 - 314 pts - 38% of space implausible

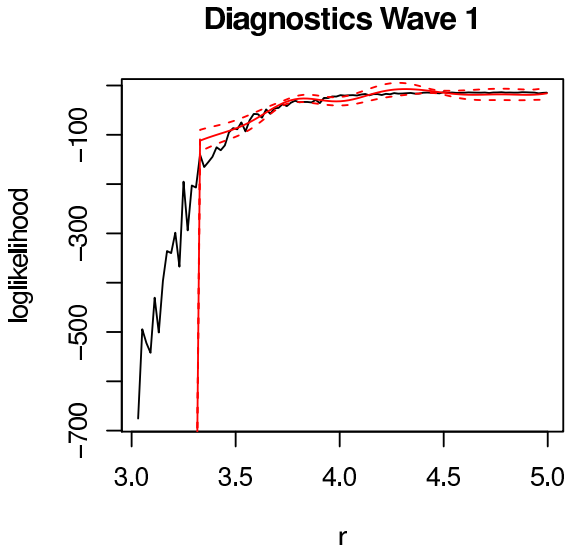


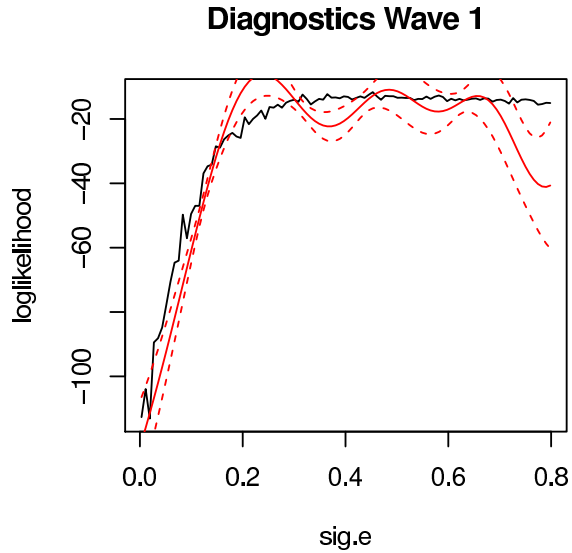


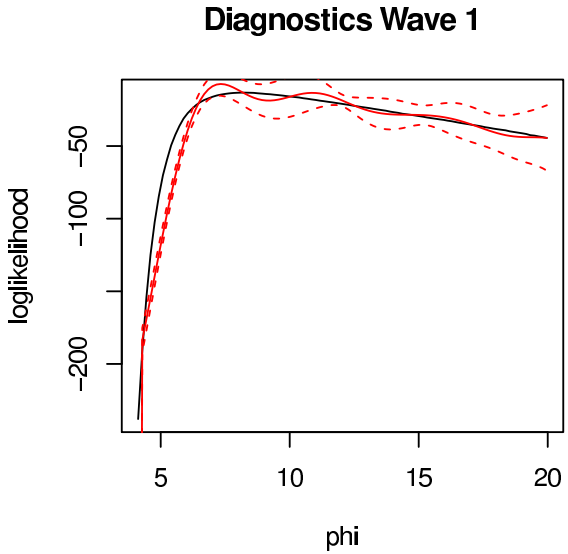


Diagnostics for GP 2 modelling log $l(\theta)$

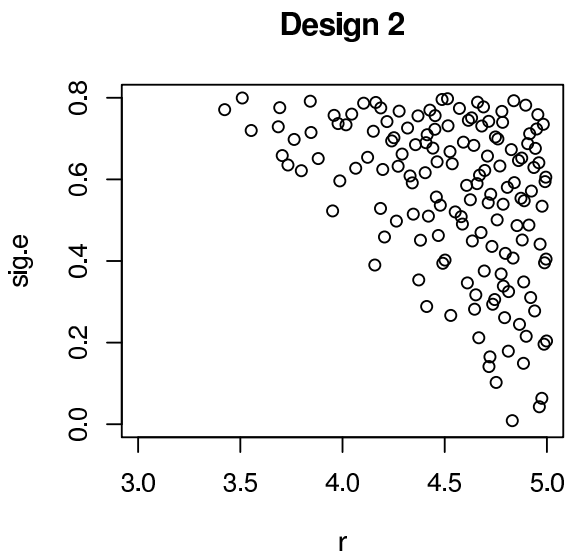
threshold = -21.8

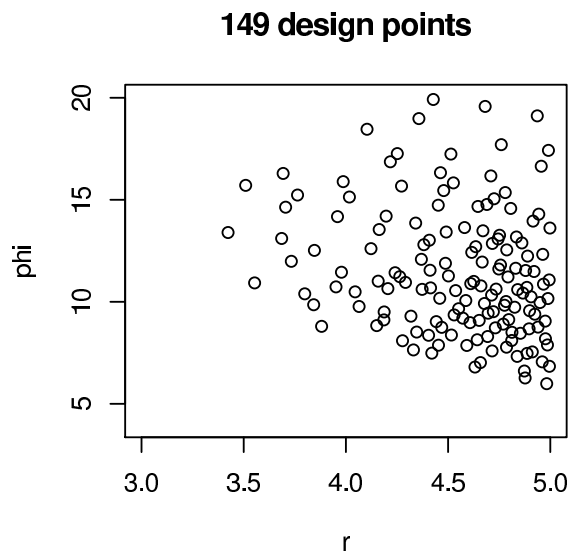


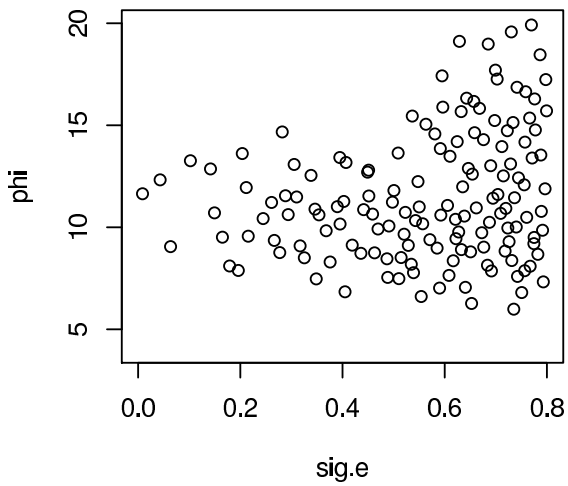




Design 3 - 149 pts - 62% of space implausible

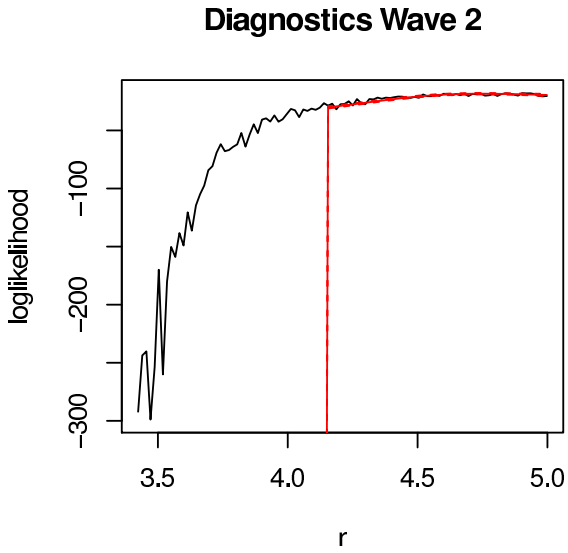


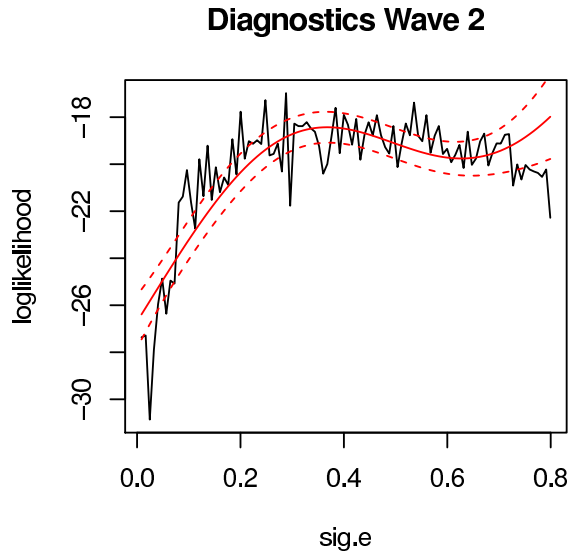


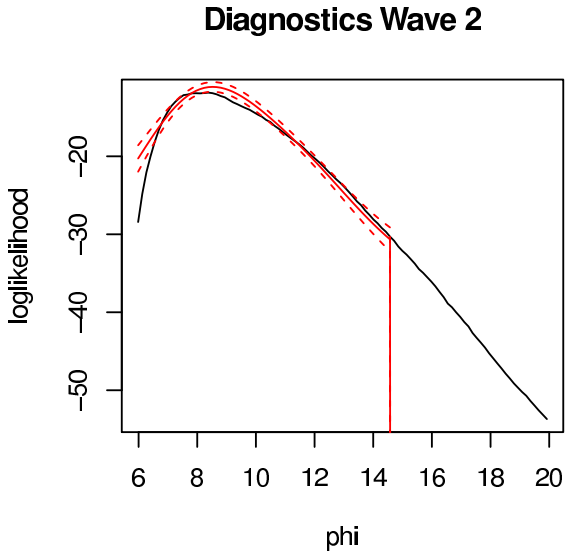


Diagnostics for GP 3 modelling $\log l(\theta)$

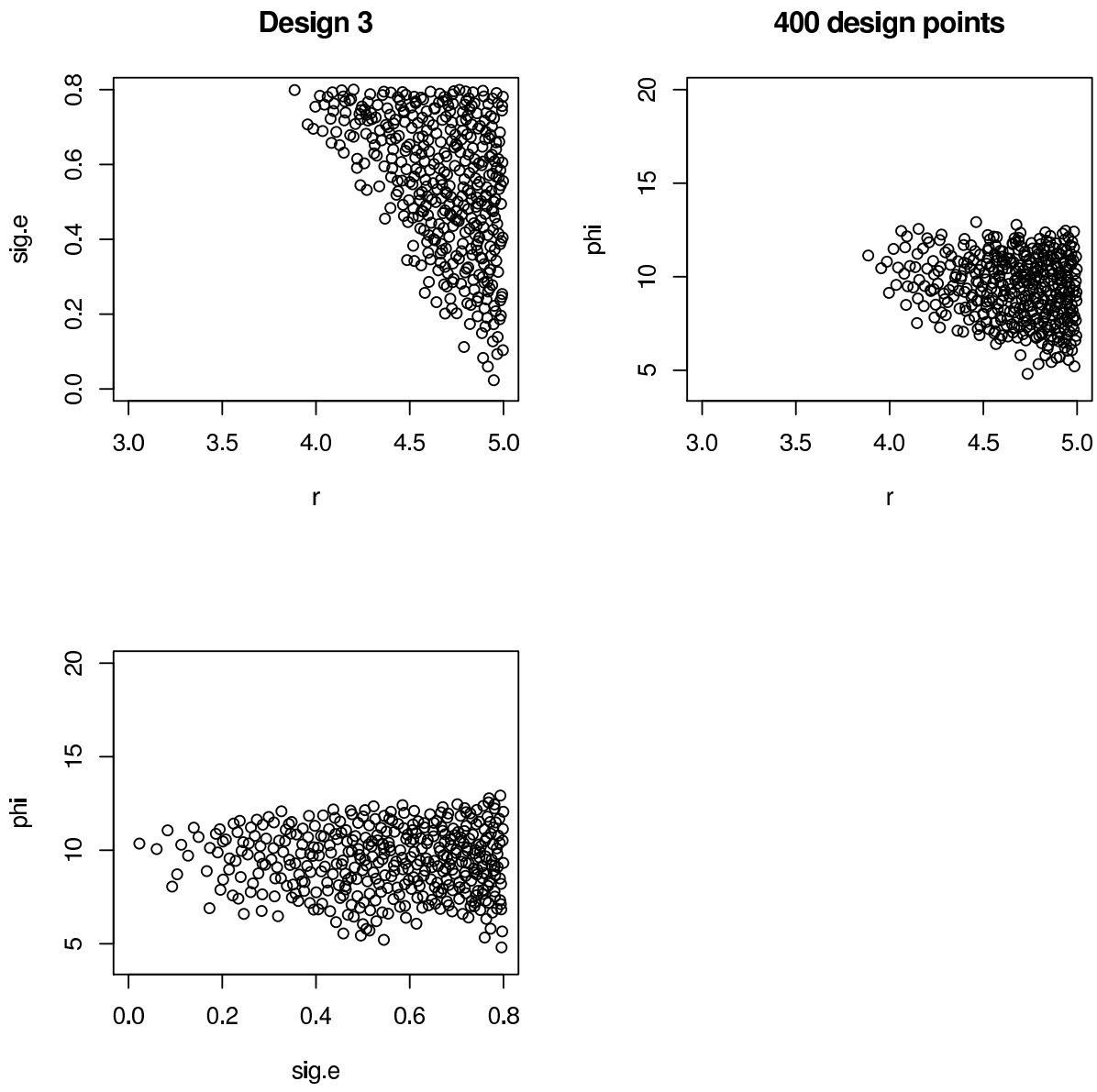
Threshold = -20.7





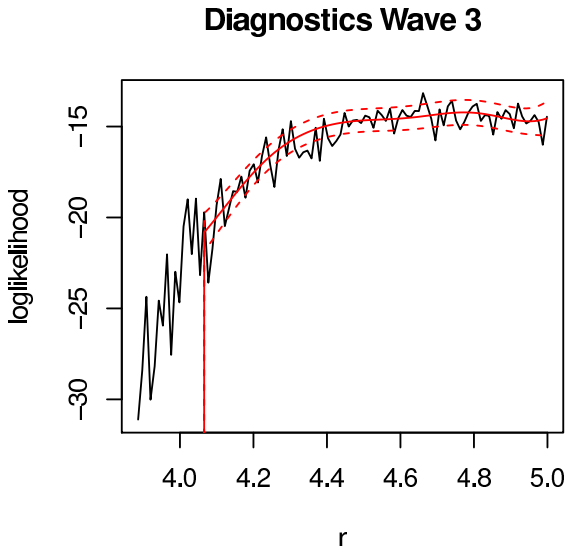


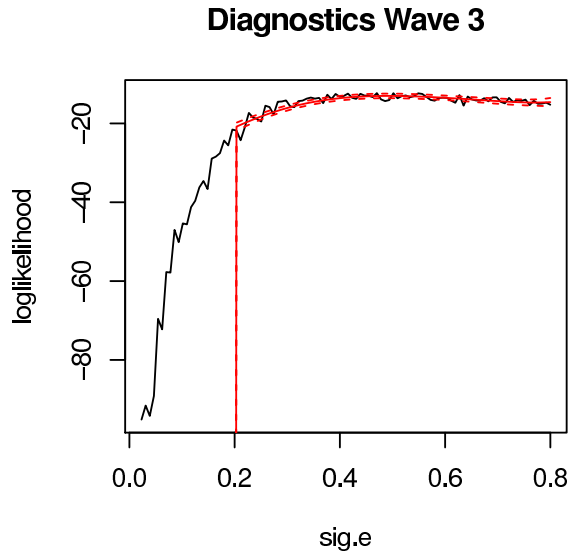
Design 4 - 400 pts - 95% of space implausible

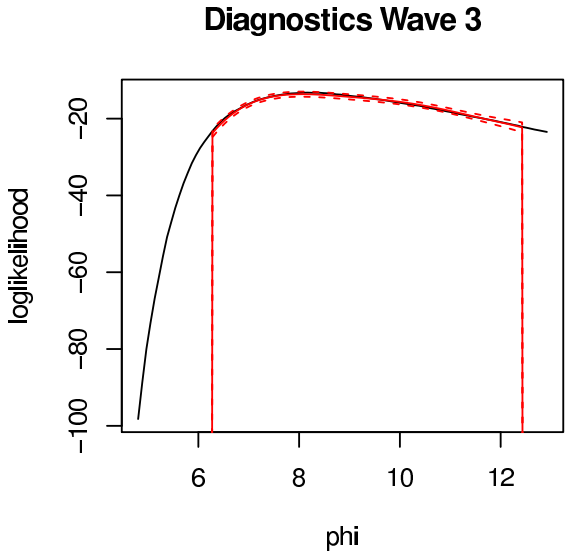


Diagnostics for GP 4 modelling log $l(\theta)$

Threshold = -16.4

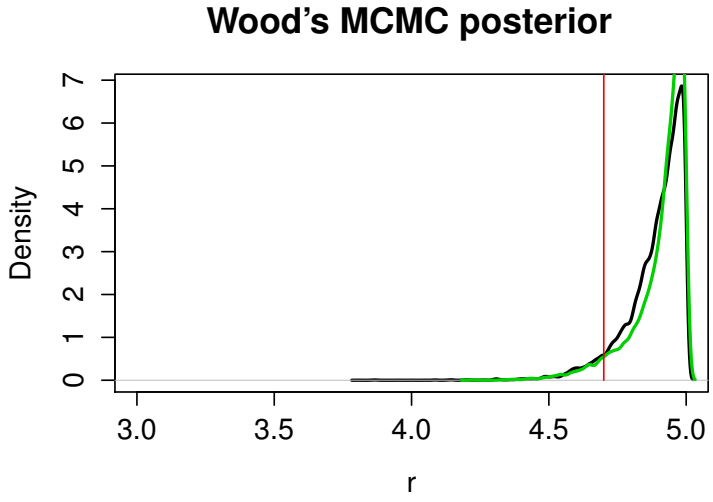


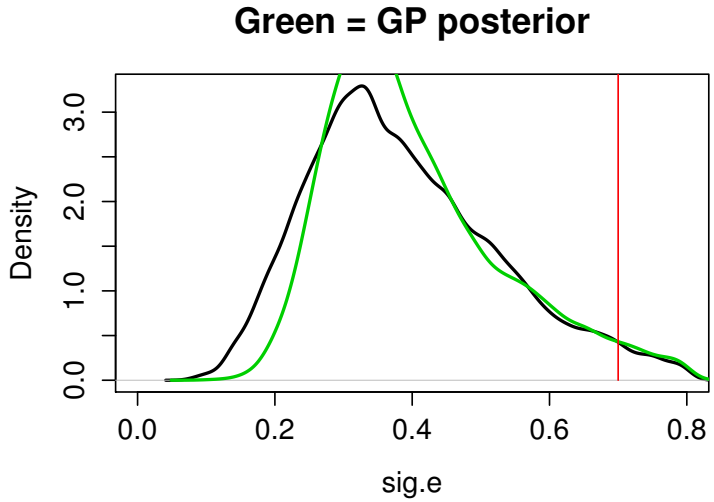


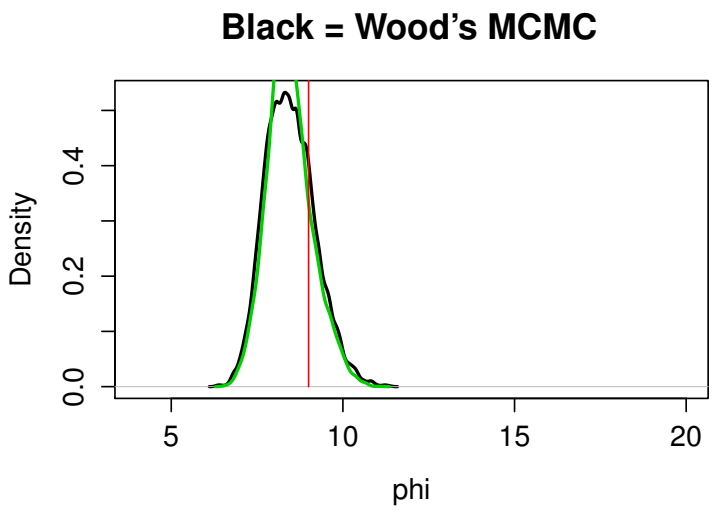


MCMC Results

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

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

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.

Instead build a sequential design $\theta_1, \theta_2, \ldots$ using our current surrogate model to guide the choice of design points according to some acquisition rule.

In practice, batch strategies are necessary if they are to be used in realistic scenarios.

- For Bayesian optimization, expected improvement is a sensible choice, i.e., to maximize $g(x)$ choose x to maximize

$$H(x) = \mathbb{E}(\max(0, g(x) - g_{max}))$$

  where $\mathbb{E}$ is with respect to the surrogate model for g.
- Gutmann and Corander use the lower confidence bound selection criterion
- Conrad *et al.* let the MCMC algorithm decide whether the current approximation is sufficiently accurate and do further simulation runs if not.

If focus is on history matching then this is a question of level set estimation...

Entropy-based acquisition

When using emulators for history-matching 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 $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$

Entropy-based acquisition

When using emulators for history-matching 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 $\mathcal{P}_\theta = \{\theta : f(\theta) \in \mathcal{P}_D\}$

- The entropy of the classification surface is

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

- We could choose the next design point where we are most uncertain.

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

  This is numerically simple, but the additional design points tend to accumulate on the edge of the plausible region

Expected average entropy

Chevalier *et al.* 2014, Holden, *et al.* 2015

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

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

- Choose θ_{n+1} to minimise the expected average entropy

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

where

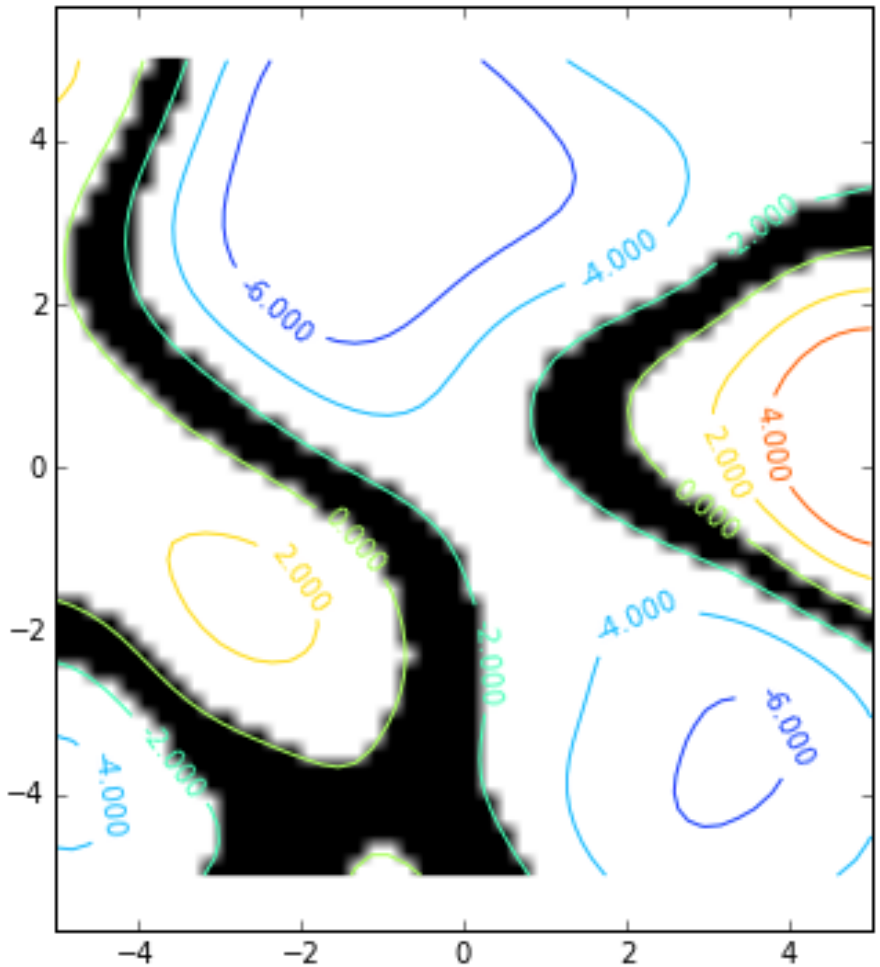
$$J_n(\theta) = \mathbb{E}(E_{n+1} | \theta_{n+1} = \theta)$$

This is computationally costly, but we can use an additional Bayesian optimization step to minimize $J(\theta)$.

History match

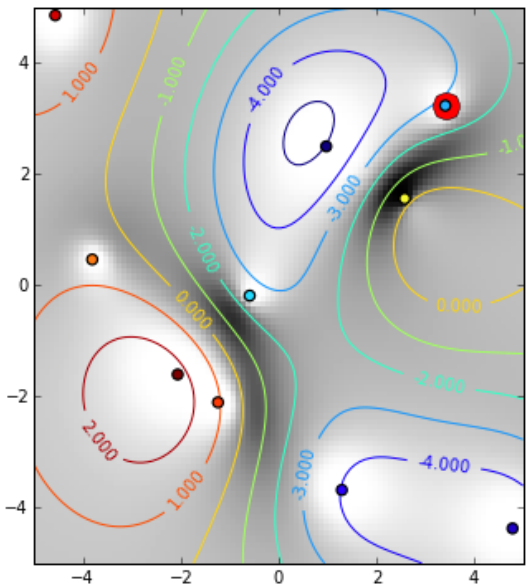
Can we learn the following plausible set?

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



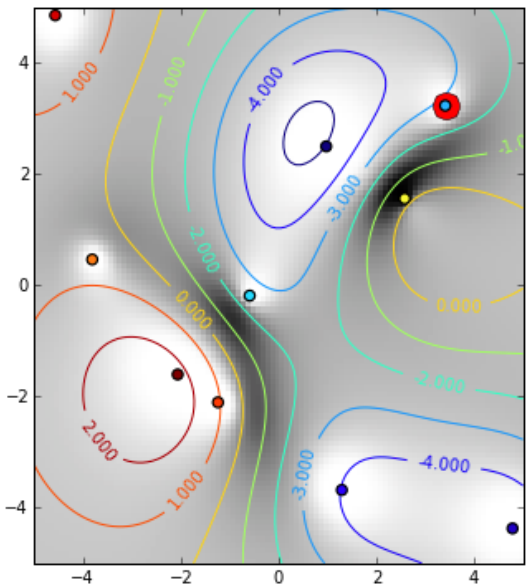
Iteration 10

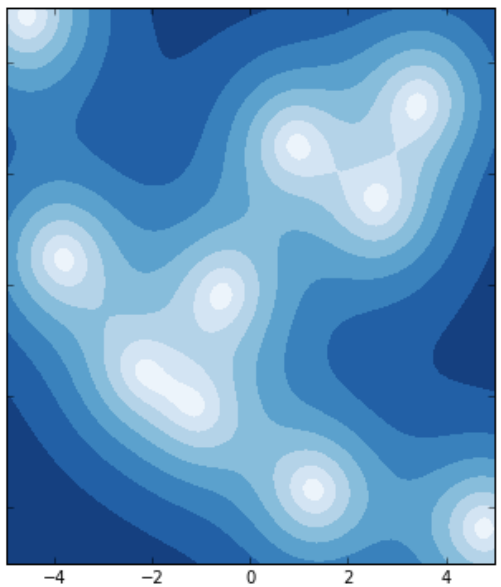
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$



Iteration 10

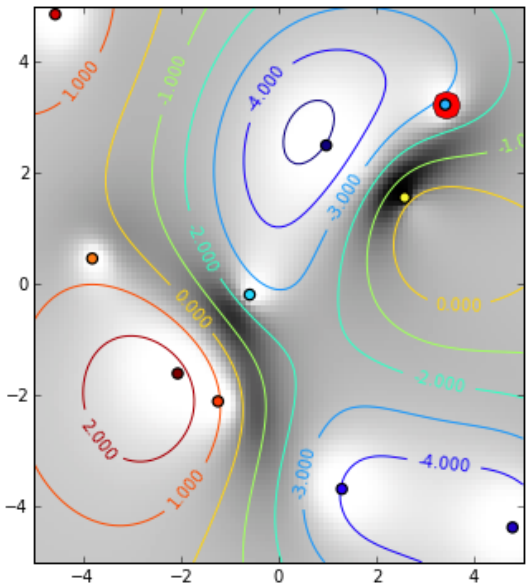
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

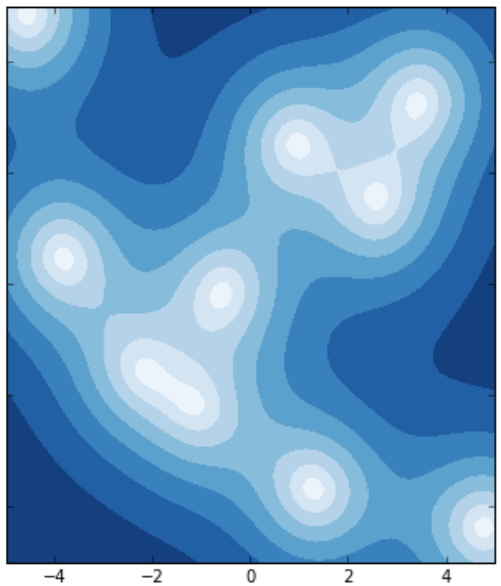


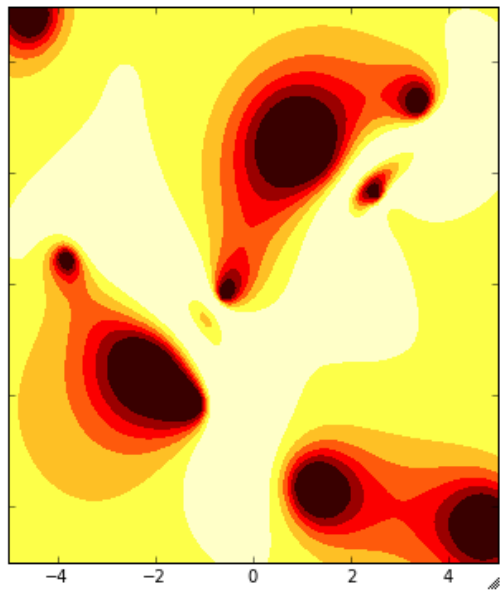


Iteration 10

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

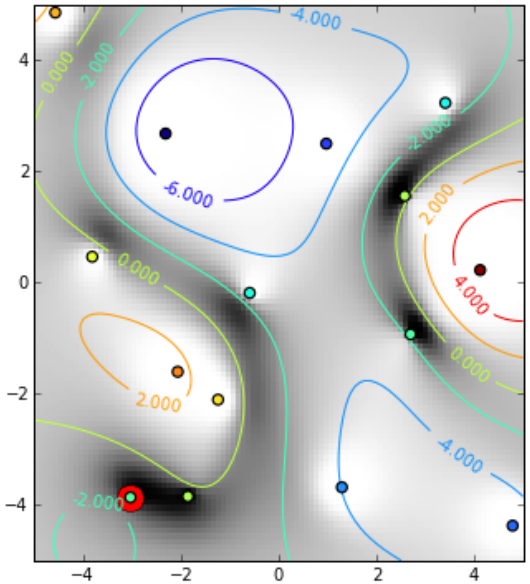


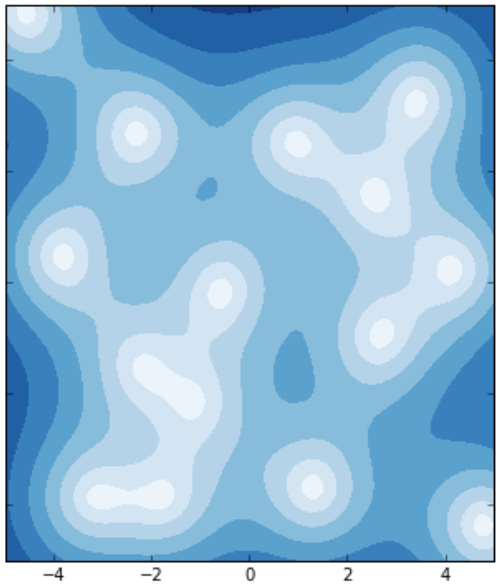


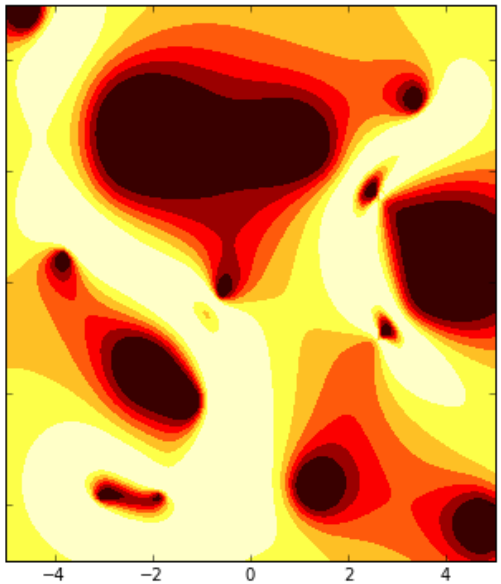


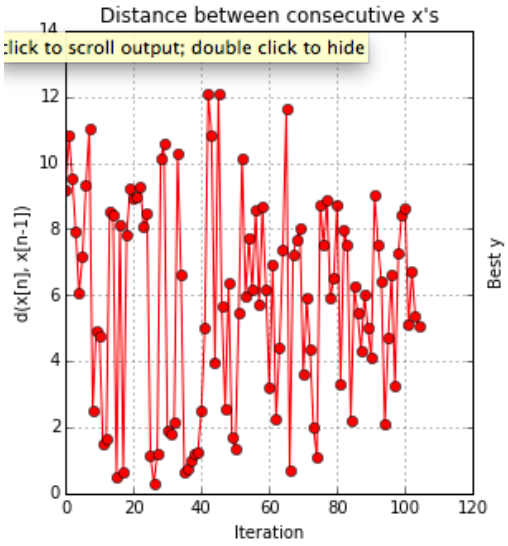
Iteration 15

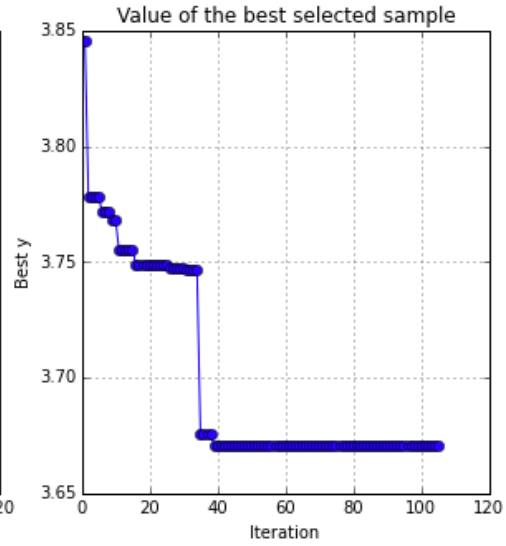
Left=$p(\theta)$, middle= $E(\theta)$, right = $\tilde{J}(\theta)$

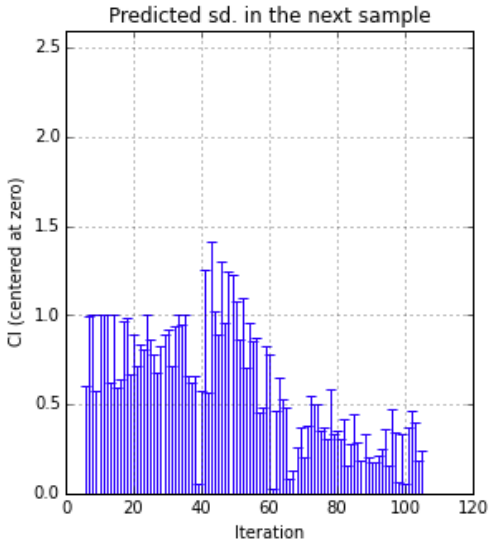




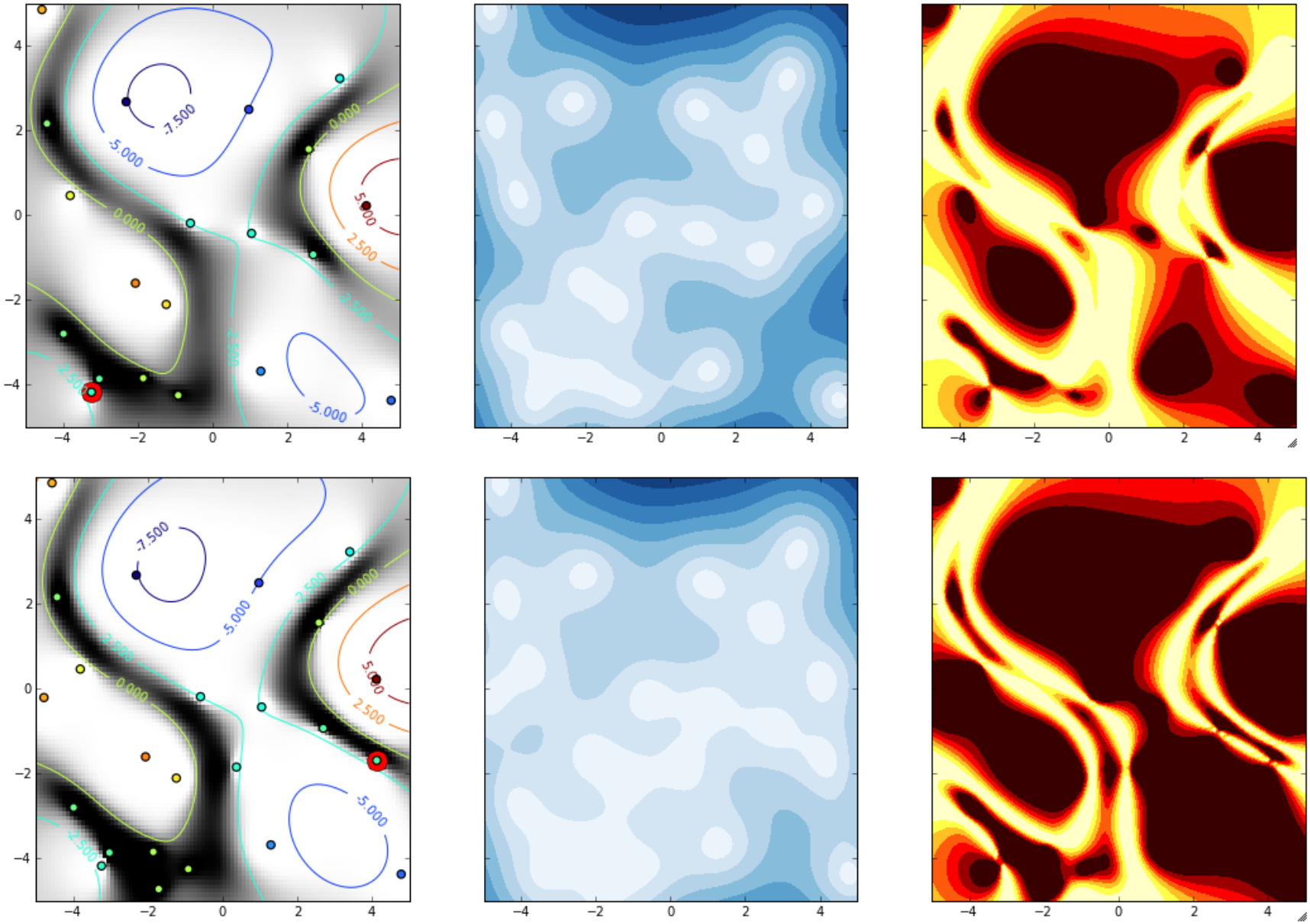






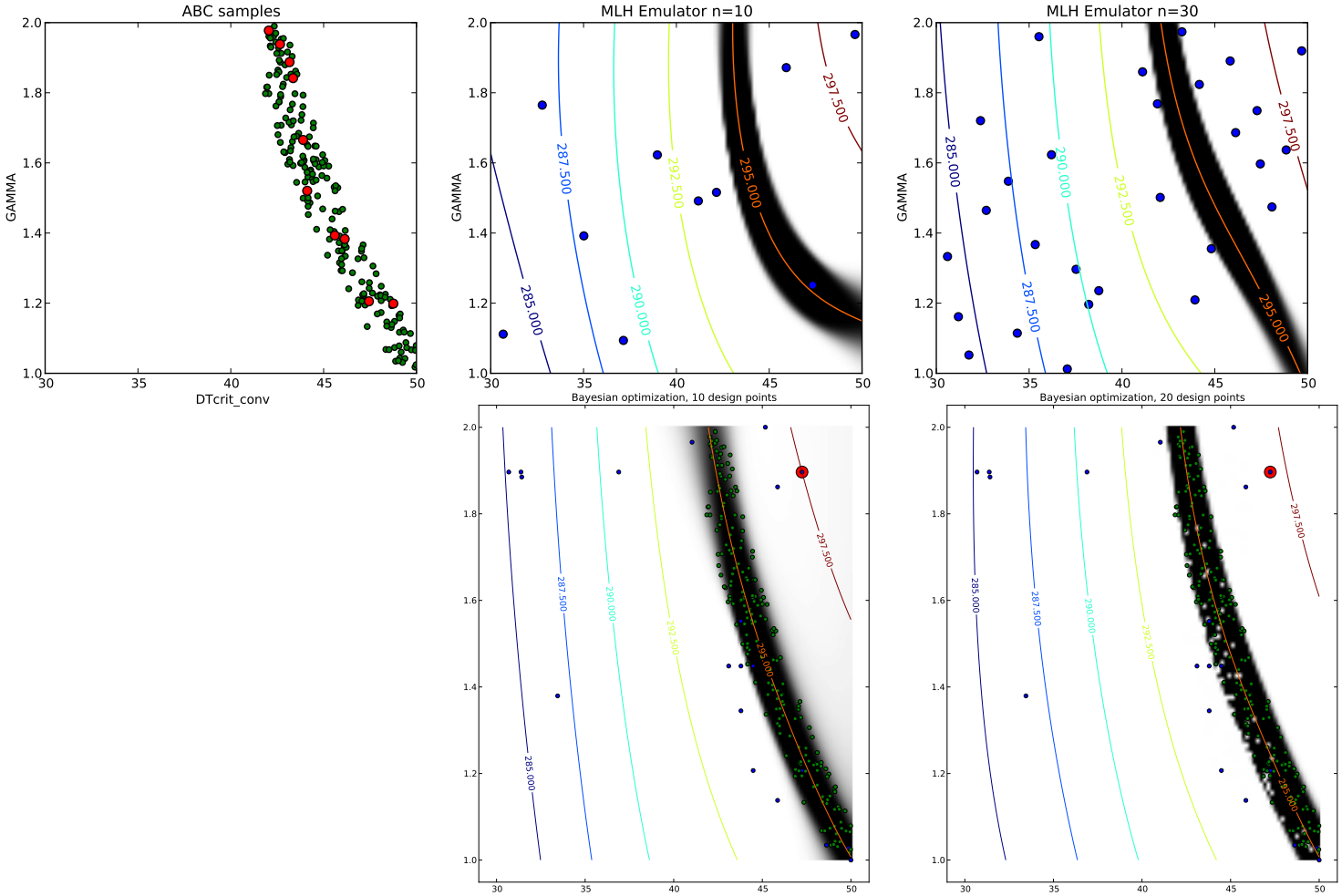


Iterations 20 and 24



Video

EPm: climate model, Holden *et al.* 2016



If we care about the posterior, what should we use?

- History matching waves with a calibration polish with a space filling design (Williamson and Vernon 2015)?

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

Variance-tolerance trade-off

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, known variance σ^2, $\mu \sim U[a, b]$.
In this case, if we use $\rho(\mathbf{D}, \mathbf{X}) = |\bar{\mathbf{D}} - \bar{\mathbf{X}}|$ in ABC with a uniform kernel, then calculation of $\pi_{ABC}(\mu)$ and $\pi(\mu|D)$ is possible.

We can show that

$$\mathbb{V}\text{ar}_{ABC}(\mu) \approx \frac{\sigma^2}{n} + \frac{1}{3}\epsilon^2$$

$$d_{TV}(\pi_{ABC}(\mu), \pi(\mu|\mathcal{D})) \approx \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2)$$

The tolerance required for a given accuracy depends on the size of the posterior variance σ^2/n.

- Can we include the model error and surrogate model variance in this calculation?

My 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: for deterministic sims quantification is hard with little methodological development. For stochastic simulators???
- Rules of thumb: how costly does the simulator need to be to make surrogate modelling worthwhile? what are good preliminary values for number of design points, number of simulator replicates etc?

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!