Introduction to ABC (Approximate Bayesian computation)

Richard Wilkinson

School of Mathematics and Statistics
University of Sheffield

September 14, 2016

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?

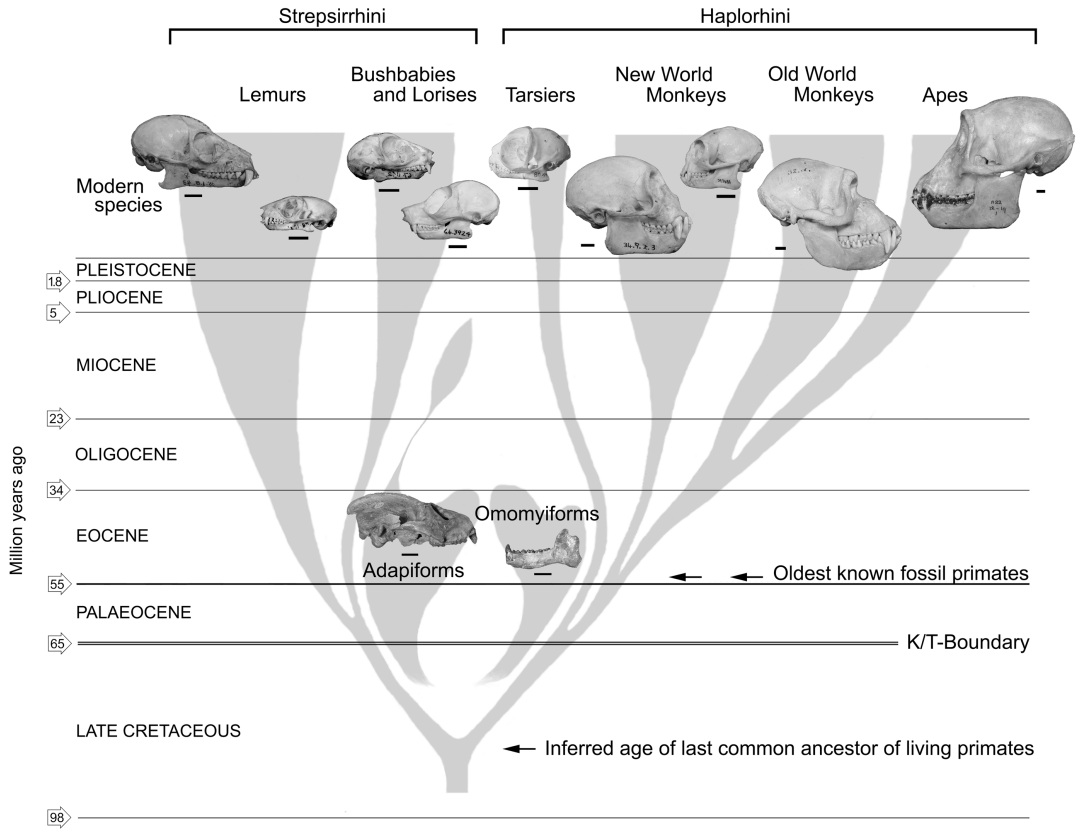
Calibration

- 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

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

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



Intractability

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

- usual intractability in Bayesian inference is not knowing $\pi(D)$.
- a problem is doubly intractable if $\pi(D|\theta) = c_\theta p(D|\theta)$ with c_θ unknown (cf Murray, Ghahramani and MacKay 2006)
- a problem is completely intractable if $\pi(D|\theta)$ is unknown and can't be evaluated (unknown is subjective). I.e., if the analytic distribution of the simulator, $f(\theta)$, run at θ is unknown.

Completely intractable models are where we need to resort to ABC methods

Approximate Bayesian Computation (ABC)

If the likelihood function is intractable, then ABC (approximate Bayesian computation) is one of the few approaches we can use to do inference.

Approximate Bayesian Computation (ABC)

If the likelihood function is intractable, then ABC (approximate Bayesian computation) is one of the few approaches we can use to do inference.

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').

Approximate Bayesian computation (ABC)

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

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

ABC methods can be crude but they have an important role to play.

Approximate Bayesian computation (ABC)

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

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

ABC methods can be crude but they have an important role to play.

First ABC paper candidates

- Beaumont *et al.* 2002
- Tavaré *et al.* 1997 or Pritchard *et al.* 1999
- Or Diggle and Gratton 1984 or Rubin 1984
- ...

Plan

i. Basics
ii. Efficient sampling algorithms
iii. Links to other approaches
iv. Regression adjustments/ post-hoc corrections
v. Expensive simulators

Basics

'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

The acceptance rate is $\int \mathbb{P}(D|\theta)\pi(\theta)\mathrm{d}\theta = \mathbb{P}(D)$.

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:

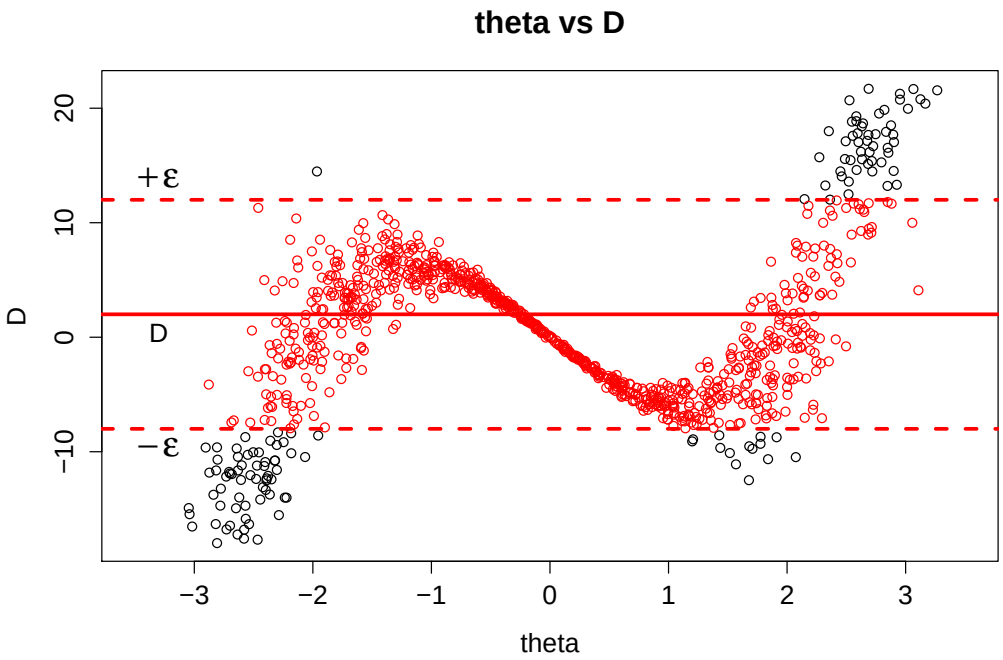
Uniform Rejection Algorithm

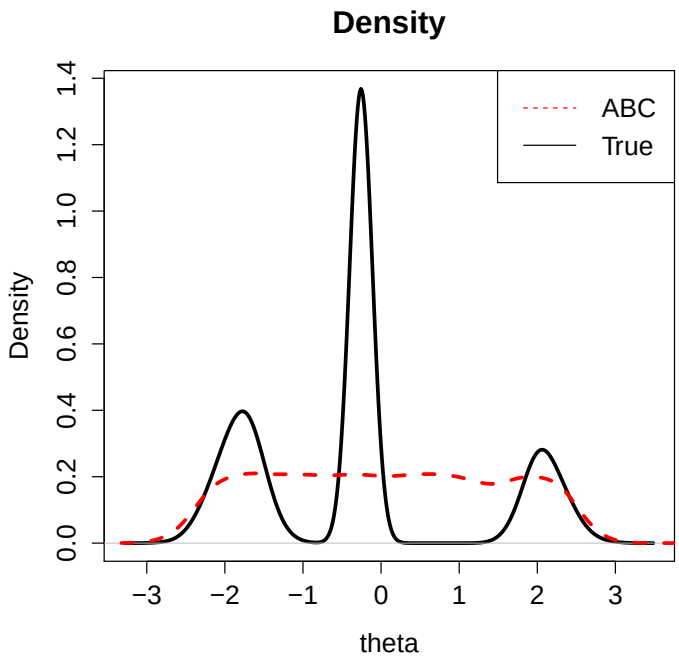
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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

$\epsilon = 10$

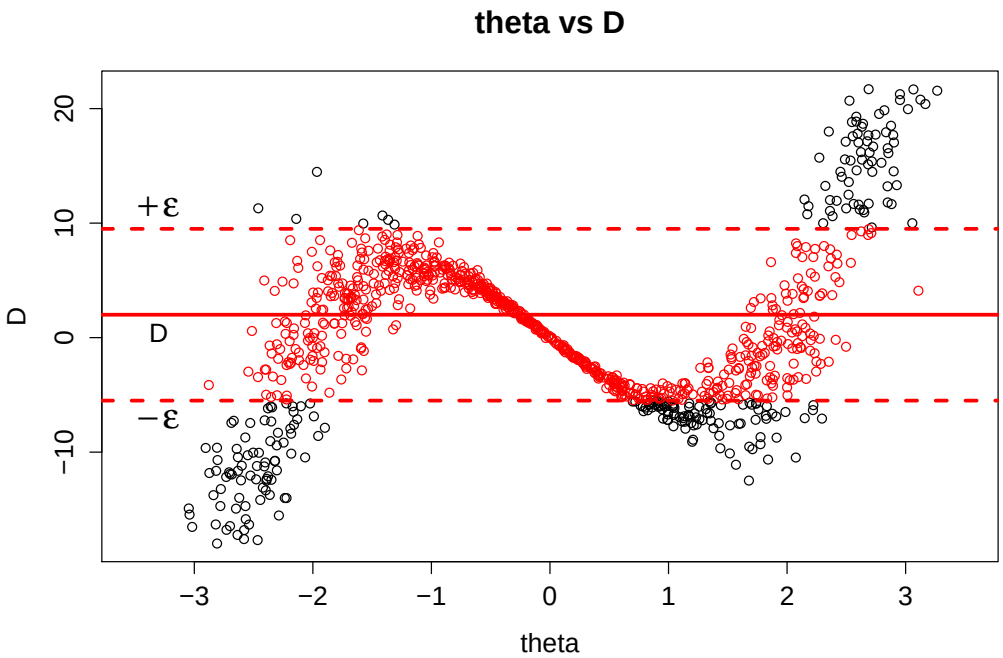


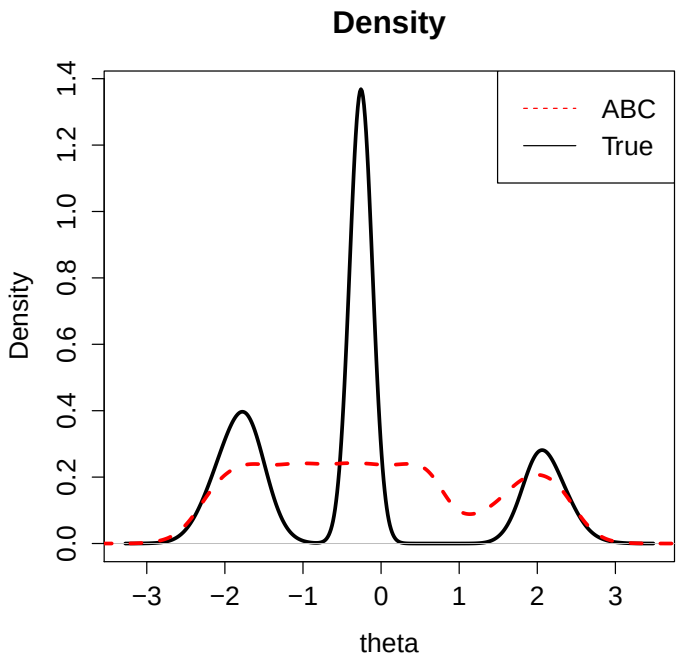


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

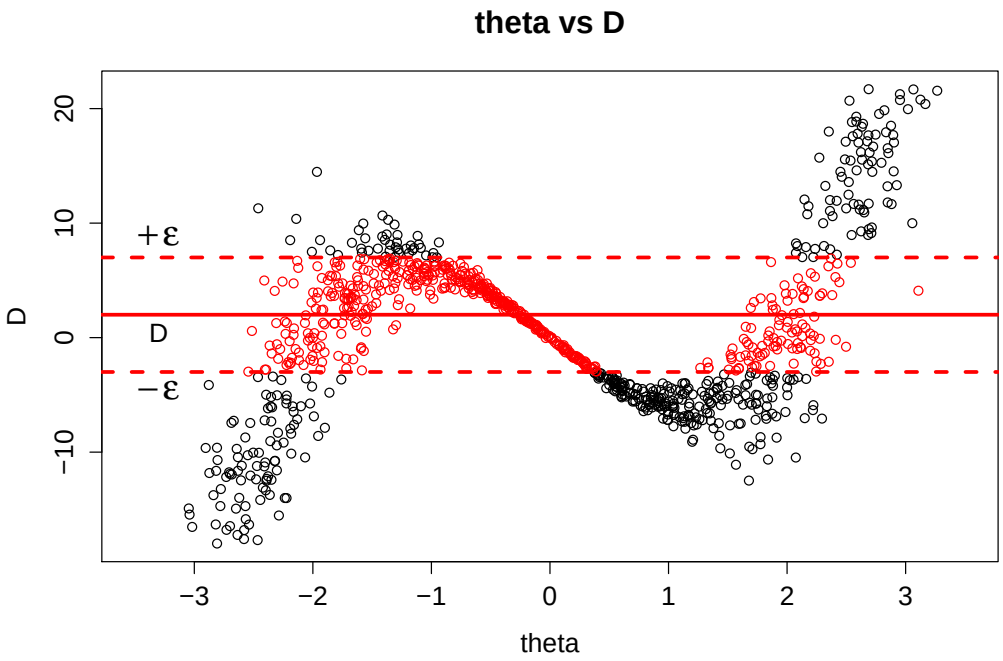
$$\rho(D, X) = |D - X|, \qquad D = 2$$

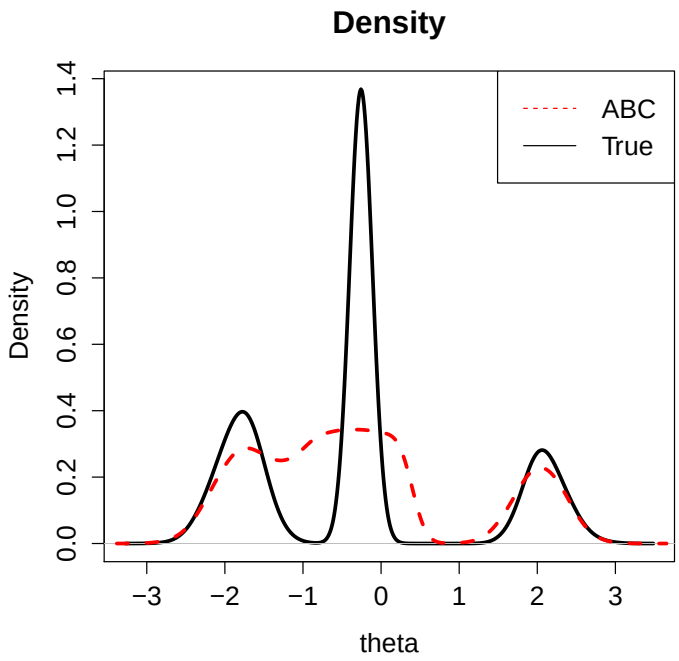
$\epsilon = 7.5$


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

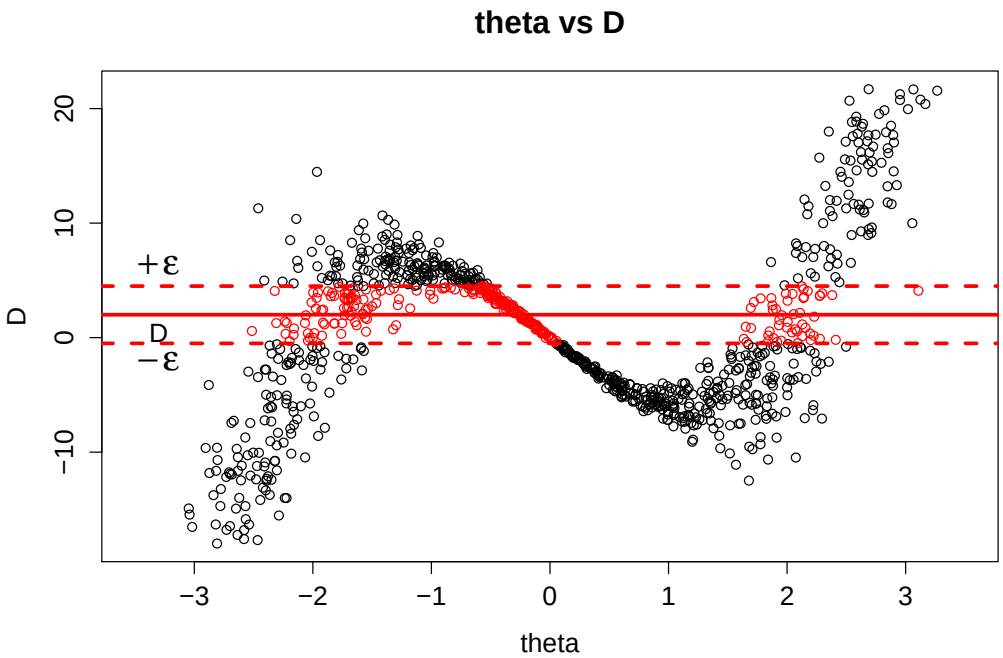

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

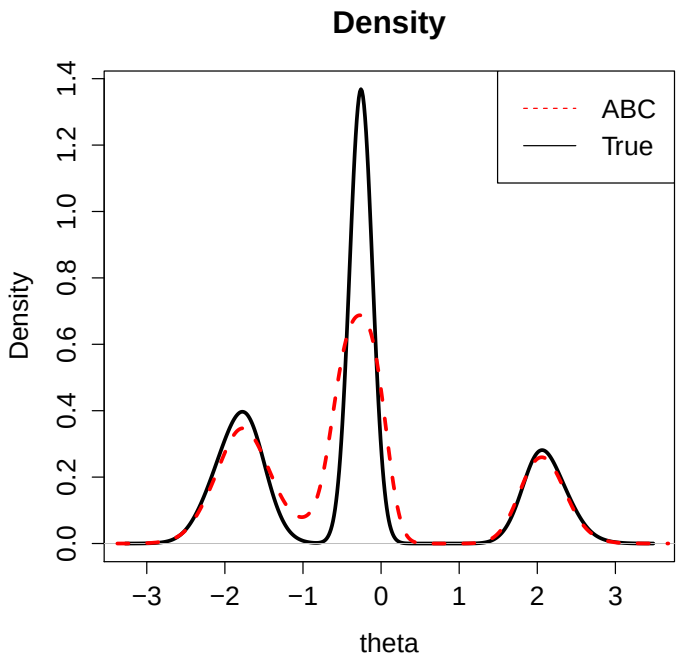
$\epsilon = 5$


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

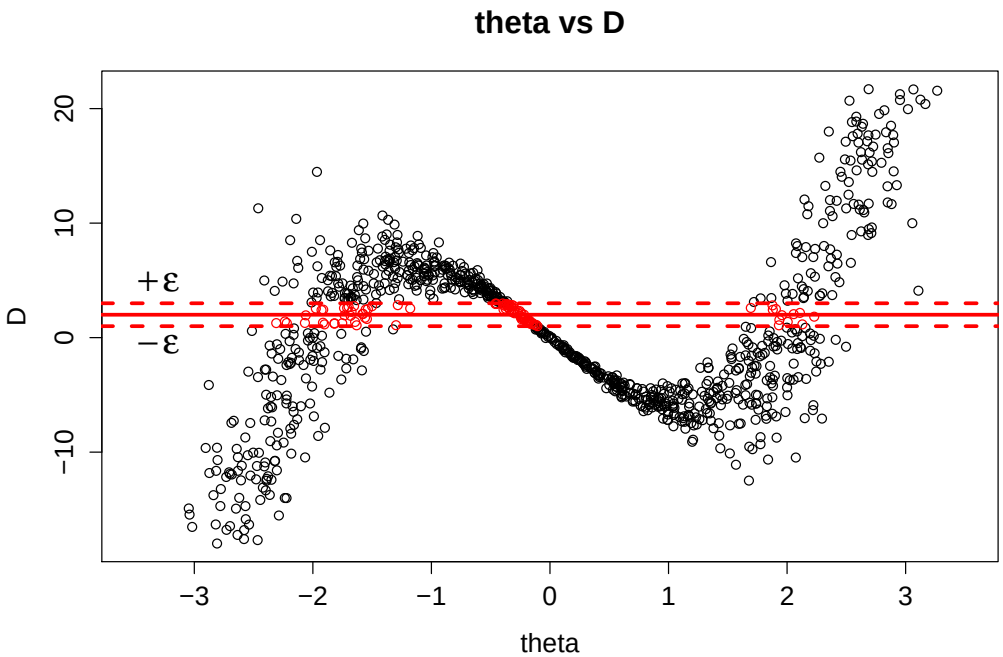

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

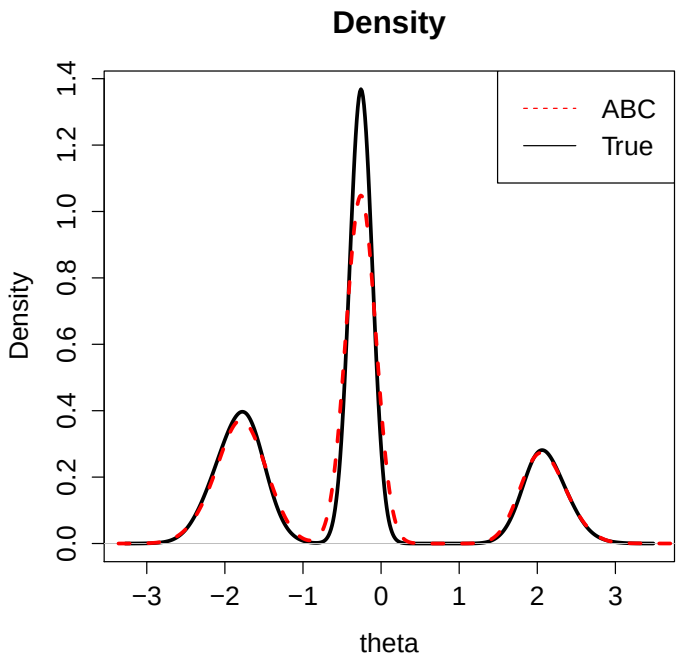
$\epsilon = 2.5$


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 field data - curse of dimensionality

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 field data - curse of dimensionality

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

ABC as a probability model

Wilkinson 2008, 2013

We wanted to solve the inverse problem

$$D = f(\theta)$$

but instead ABC solves

$$D = f(\theta) + e.$$

ABC as a probability model

Wilkinson 2008, 2013

We wanted to solve the inverse problem

$$D = f(\theta)$$

but instead ABC solves

$$D = f(\theta) + e.$$

ABC gives 'exact' inference under a different model!

We can show that

Proposition

If $\rho(D, X) = |D - X|$, then ABC samples from the posterior distribution of θ given D where we assume $D = f(\theta) + e$ and that

$$e \sim U[-\epsilon, \epsilon]$$

Generalized ABC (GABC)

Generalized rejection ABC (Rej-GABC)

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$
2. Accept (θ, X) if $U \sim U[0,1] \leq \frac{\pi_\epsilon(D|X)}{\max_x \pi_\epsilon(D|x)}$

In uniform ABC we take

$$\pi_\epsilon(D|X) = \begin{cases} 1 & \text{if } \rho(D,X) \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

which recovers the *uniform* ABC algorithm.

2' Accept θ ifF $\rho(D,X) \leq \epsilon$

Generalized ABC (GABC)

Generalized rejection ABC (Rej-GABC)

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$
2. Accept (θ, X) if $U \sim U[0,1] \leq \frac{\pi_\epsilon(D|X)}{\max_x \pi_\epsilon(D|x)}$

In uniform ABC we take

$$\pi_\epsilon(D|X) = \begin{cases} 1 & \text{if } \rho(D,X) \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

which recovers the *uniform* ABC algorithm.

2' Accept θ ifF $\rho(D,X) \leq \epsilon$

We can use $\pi_\epsilon(D|x)$ to describe the relationship between the simulator and reality, e.g., measurement error and simulator discrepancy.

- We don't need to assume uniform error!

Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'

- Summary statistic $S(D)$ - controls 'information loss'

Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ and hence find good approximations
  - constrained by limitations on how much computation we can do - rules out expensive simulators
  - how do we relate simulators to reality
- Summary statistic $S(D)$ - controls 'information loss'

Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ and hence find good approximations
  - constrained by limitations on how much computation we can do - rules out expensive simulators
  - how do we relate simulators to reality
- Summary statistic $S(D)$ - controls 'information loss'

Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ and hence find good approximations
  - constrained by limitations on how much computation we can do - rules out expensive simulators
  - how do we relate simulators to reality
- Summary statistic $S(D)$ - controls 'information loss'
  - inference is based on $\pi(\theta|S(D))$ rather than $\pi(\theta|D)$
  - a combination of expert judgement, and stats/ML tools can be used to find informative summaries

Efficient Algorithms

References:

- Marjoram *et al.* 2003
- Sisson *et al.* 2007
- Beaumont *et al.* 2008
- Toni *et al.* 2009
- Del Moral *et al.* 2011
- Drovandi *et al.* 2011

ABCifying Monte Carlo methods

Rejection ABC is the basic ABC algorithm

- Inefficient as it repeatedly samples from prior

More efficient sampling algorithms allow us to make better use of the available computational resource: spend more time in regions of parameter space likely to lead to accepted values.

- allows us to use smaller values of ϵ, and hence finding better approximations

Most Monte Carlo algorithms now have ABC versions for when we don't know the likelihood: IS, MCMC, SMC ($\times n$), EM, EP etc

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012

We are targeting the joint distribution

$$\pi_{ABC}(\theta, x|D) \propto \pi_{\epsilon}(D|x)\pi(x|\theta)\pi(\theta)$$

To explore the (θ, x) space, proposals of the form

$$Q((\theta, x), (\theta', x')) = q(\theta, \theta')\pi(x'|\theta')$$

seem to be inevitable.

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012

We are targeting the joint distribution

$$\pi_{ABC}(\theta, x|D) \propto \pi_{\epsilon}(D|x)\pi(x|\theta)\pi(\theta)$$

To explore the (θ, x) space, proposals of the form

$$Q((\theta, x), (\theta', x')) = q(\theta, \theta')\pi(x'|\theta')$$

seem to be inevitable.

The Metropolis-Hastings (MH) acceptance probability is then

$$r = \frac{\pi_{ABC}(\theta', x'|D)Q((\theta', x'), (\theta, x))}{\pi_{ABC}(\theta, x|D)Q((\theta, x), (\theta', x'))}$$

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012

We are targeting the joint distribution

$$\pi_{ABC}(\theta, x|D) \propto \pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)$$

To explore the (θ, x) space, proposals of the form

$$Q((\theta, x), (\theta', x')) = q(\theta, \theta')\pi(x'|\theta')$$

seem to be inevitable.

The Metropolis-Hastings (MH) acceptance probability is then

$$\begin{aligned} r &= \frac{\pi_{ABC}(\theta', x'|D)Q((\theta', x'), (\theta, x))}{\pi_{ABC}(\theta, x|D)Q((\theta, x), (\theta', x'))} \\ &= \frac{\pi_\epsilon(D|x')\pi(x'|\theta')\pi(\theta')q(\theta', \theta)\pi(x|\theta)}{\pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)q(\theta, \theta')\pi(x'|\theta')} \end{aligned}$$

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012

We are targeting the joint distribution

$$\pi_{ABC}(\theta, x|D) \propto \pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)$$

To explore the (θ, x) space, proposals of the form

$$Q((\theta, x), (\theta', x')) = q(\theta, \theta')\pi(x'|\theta')$$

seem to be inevitable.

The Metropolis-Hastings (MH) acceptance probability is then

$$\begin{aligned} r &= \frac{\pi_{ABC}(\theta', x'|D)Q((\theta', x'), (\theta, x))}{\pi_{ABC}(\theta, x|D)Q((\theta, x), (\theta', x'))} \\ &= \frac{\pi_\epsilon(D|x')\pi(x'|\theta')\pi(\theta')q(\theta', \theta)\pi(x|\theta)}{\pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)q(\theta, \theta')\pi(x'|\theta')} \\ &= \frac{\pi_\epsilon(D|x')q(\theta', \theta)\pi(\theta')}{\pi_\epsilon(D|x)q(\theta, \theta')\pi(\theta)} \end{aligned}$$

This gives the following MCMC algorithm

MH-ABC - $P_{Marj}(\theta_0, \cdot)$

1. Propose a move from $z_t = (\theta, x)$ to (θ', x') using proposal Q above.
2. Accept move with probability

$$r((\theta, x), (\theta', x')) = \min\left(1, \frac{\pi_\epsilon(D|x')q(\theta', \theta)\pi(\theta')}{\pi_\epsilon(D|x)q(\theta, \theta')\pi(\theta)}\right),$$

   otherwise set $z_{t+1} = z_t$.

This gives the following MCMC algorithm

MH-ABC - $P_{Marj}(\theta_0, \cdot)$

1. Propose a move from $z_t = (\theta, x)$ to (θ', x') using proposal Q above.
2. Accept move with probability

$$r((\theta, x), (\theta', x')) = \min\left(1, \frac{\pi_\epsilon(D|x')q(\theta', \theta)\pi(\theta')}{\pi_\epsilon(D|x)q(\theta, \theta')\pi(\theta)}\right),$$

   otherwise set $z_{t+1} = z_t$.

In practice, this algorithm often gets stuck, as the probability of generating x' near D can be tiny if ϵ is small.

Lee 2012 introduced several alternative MCMC kernels that are variance bounding and geometrically ergodic.

Sequential ABC algorithms

Sisson *et al.* 2007, Toni *et al.* 2008, Beaumont *et al.* 2009, Del Moral *et al.* 2011, Drovandi *et al.* 2011, ...

The most popular efficient ABC algorithms are those based on sequential methods.

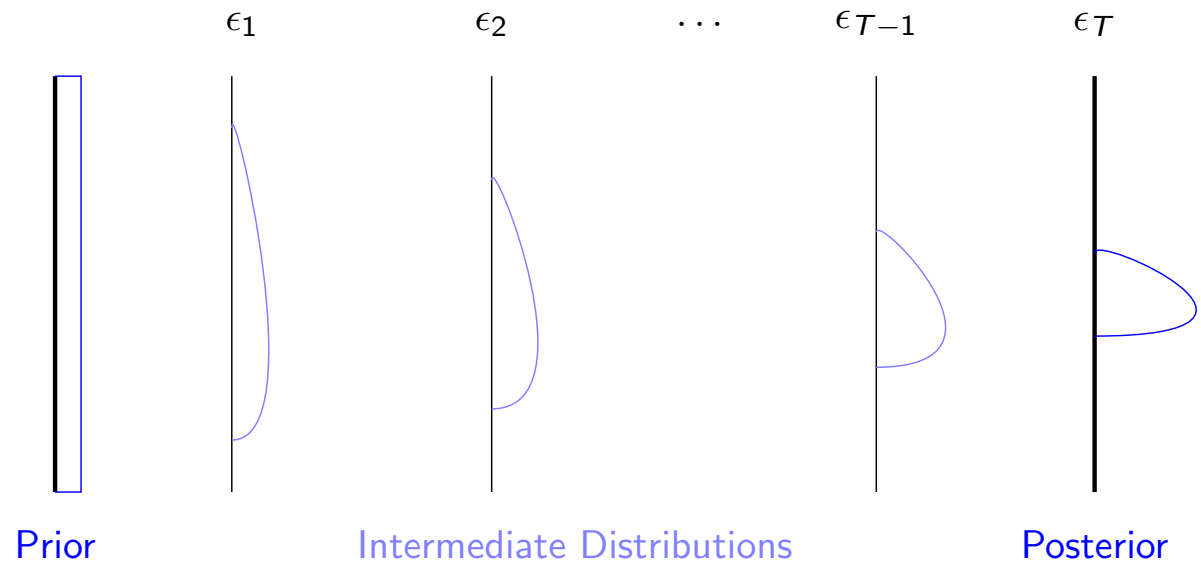
We aim to sample N particles successively from a sequence of distributions

$$\pi_1(\theta),\ \ldots,\ \pi_T(\theta) = \text{target}$$

For ABC we decide upon a sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ and let π_t be the ABC distribution found by the ABC algorithm when we use tolerance ϵ_t.

Specifically, define a sequence of target distributions

$$\pi_t(\theta, x) = \frac{\mathbb{I}_{\rho(D,x)<\epsilon_t}\pi(x|\theta)\pi(\theta)}{C_t} = \frac{\gamma_t(\theta, x)}{C_t}$$

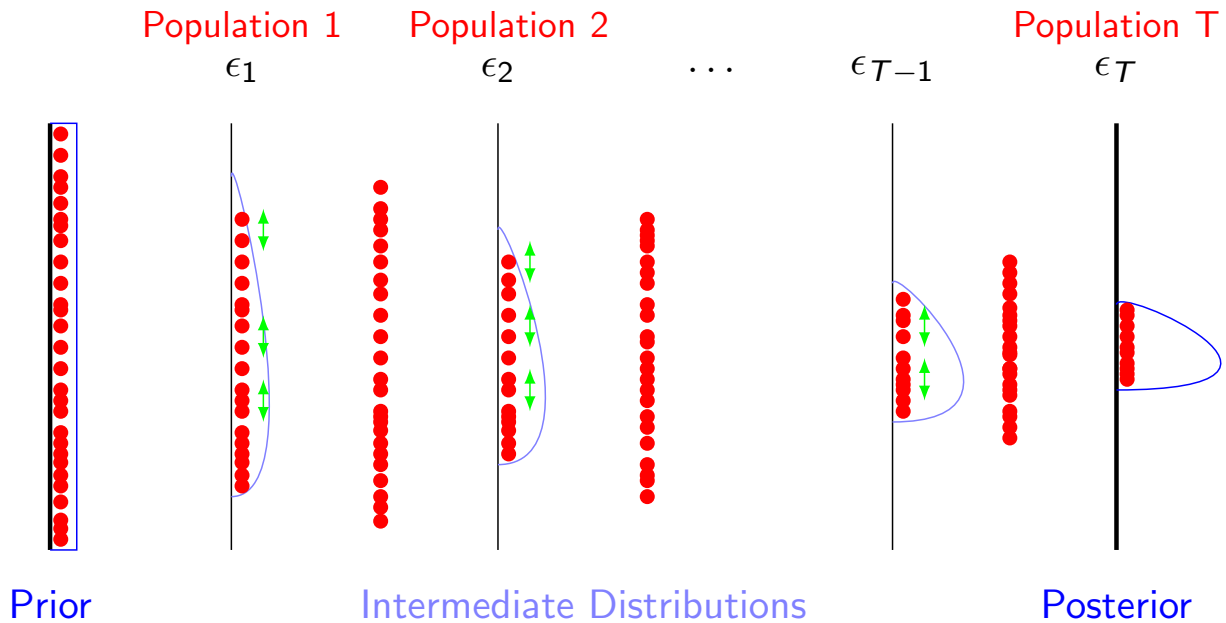


Picture from Toni and Stumpf 2010 tutorial

At each stage t, we aim to construct a weighted sample of particles that approximates $\pi_t(\theta, x)$.

$$\left\{\left(z_t^{(i)}, W_t^{(i)}\right)\right\}_{i=1}^{N} \text{ such that } \pi_t(z) \approx \sum W_t^{(i)} \delta_{z_t^{(i)}}(\mathrm{dz})$$

where $z_t^{(i)} = (\theta_t^{(i)}, x_t^{(i)})$.



Picture from Toni and Stumpf 2010 tutorial

Synthetic likelihood

The synthetic likelihood approach of Wood 2010 is an ABC algorithm which uses a Gaussian likelihood. However, instead of using

$$\pi_\epsilon(D|X) = \mathcal{N}(D; X, \epsilon)$$

and

$$\pi_{ABC}(D|\theta) = \int \mathcal{N}(D; X, \epsilon)\pi(X|\theta)dX$$

they repeatedly run the simulator at θ, generating $X_1, \ldots, X_n$, and then use

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

where μ_θ and Σ_θ is the sample mean and covariance of the (summary of the) simulator output.

Regression Adjustment

References:

- Beaumont *et al.* 2003
- Blum and Francois 2010
- Blum 2010
- Leuenberger and Wegmann 2010

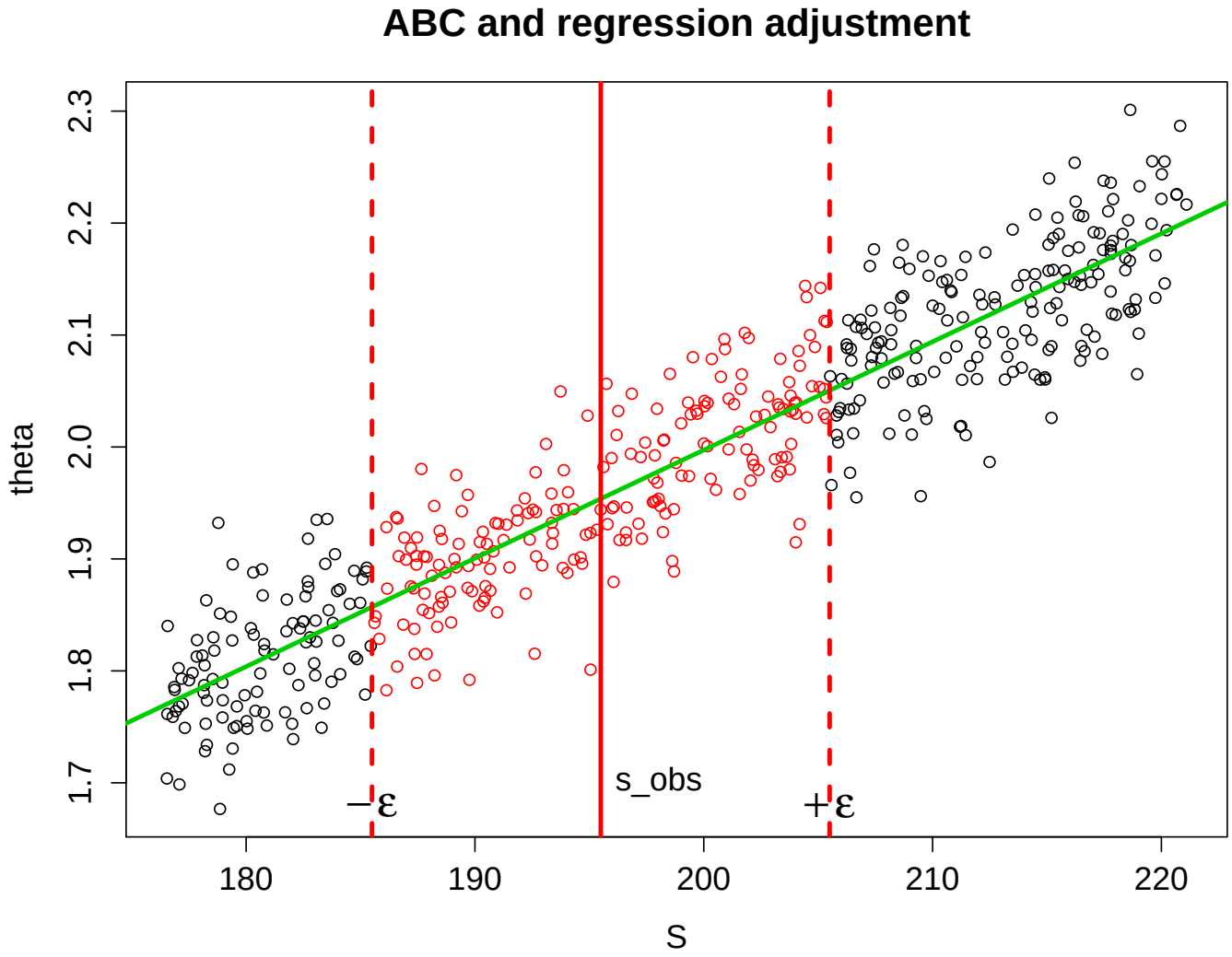
Regression Adjustment

An alternative to rejection-ABC, proposed by Beaumont *et al.* 2002, uses post-hoc adjustment of the parameter values to try to weaken the effect of the discrepancy between s and s_{obs}.

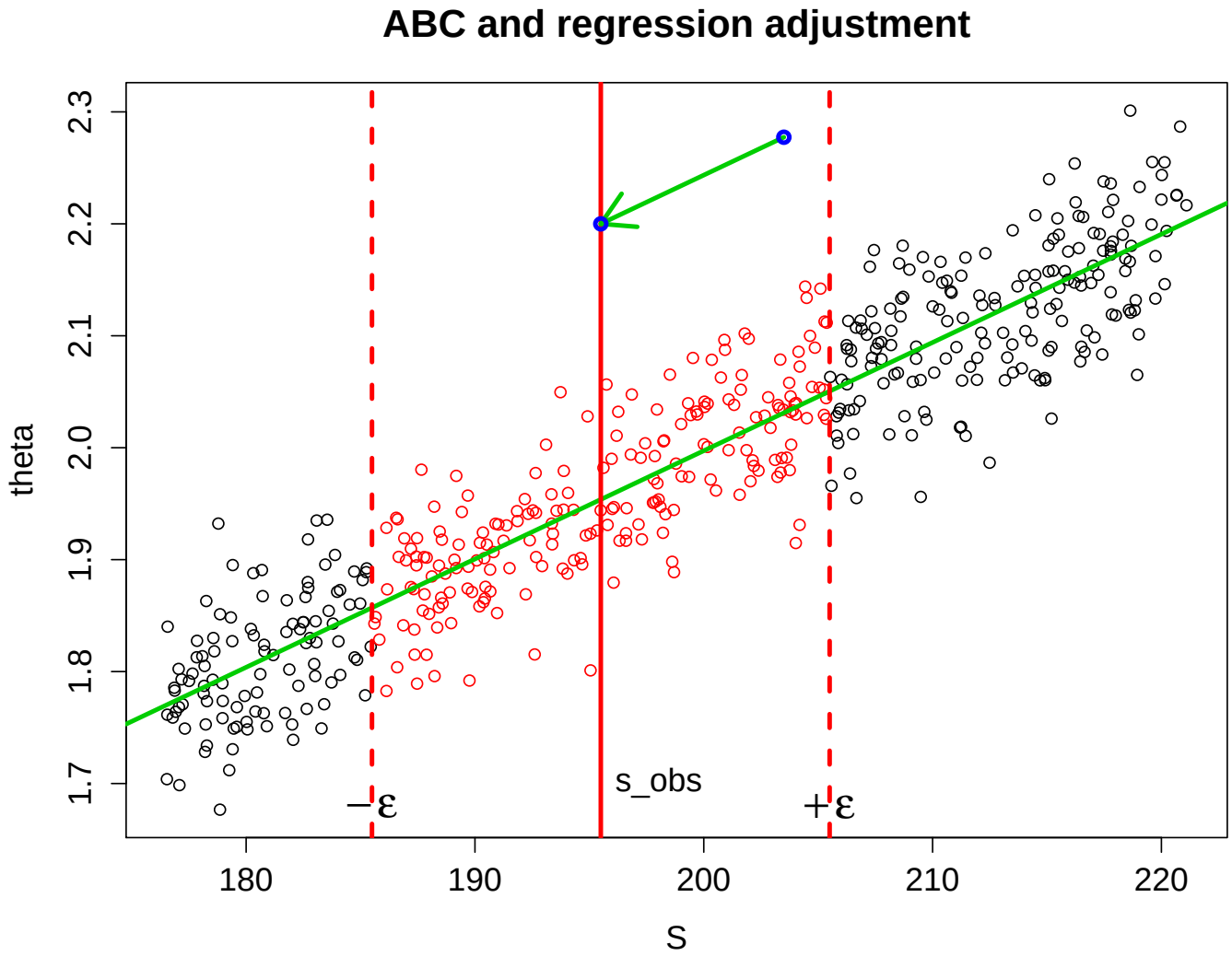
Two key ideas

- use non-parametric kernel density estimation to emphasise the best simulations
- learn a non-linear model for the conditional expectation $\mathbb{E}(\theta|s)$ as a function of s and use this to learn the posterior at s_{obs}.

These methods allow us to use a larger tolerance values and can substantially improve posterior accuracy with less computation. However, sequential algorithms can not easily be adapted, and so these methods tend to be used with simple rejection sampling.



In rejection ABC, the red points are used to approximate the histogram.



In rejection ABC, the red points are used to approximate the histogram. Using regression-adjustment, we use the estimate of the posterior mean at s_{obs} and the residuals from the fitted line to form the posterior.

Models

Beaumont *et al.* 2003 used a local linear model for $m(s)$ in the vicinity of s_{obs}

$$m(s_i) = \alpha + \beta^T s_i$$

fit by minimising

$$\sum (\theta_i - m(s_i))^2 K_\epsilon (s_i - s_{obs})$$

so that observations nearest to s_{obs} are given more weight in the fit.

Models

Beaumont *et al.* 2003 used a local linear model for $m(s)$ in the vicinity of s_{obs}

$$m(s_i) = \alpha + \beta^T s_i$$

fit by minimising

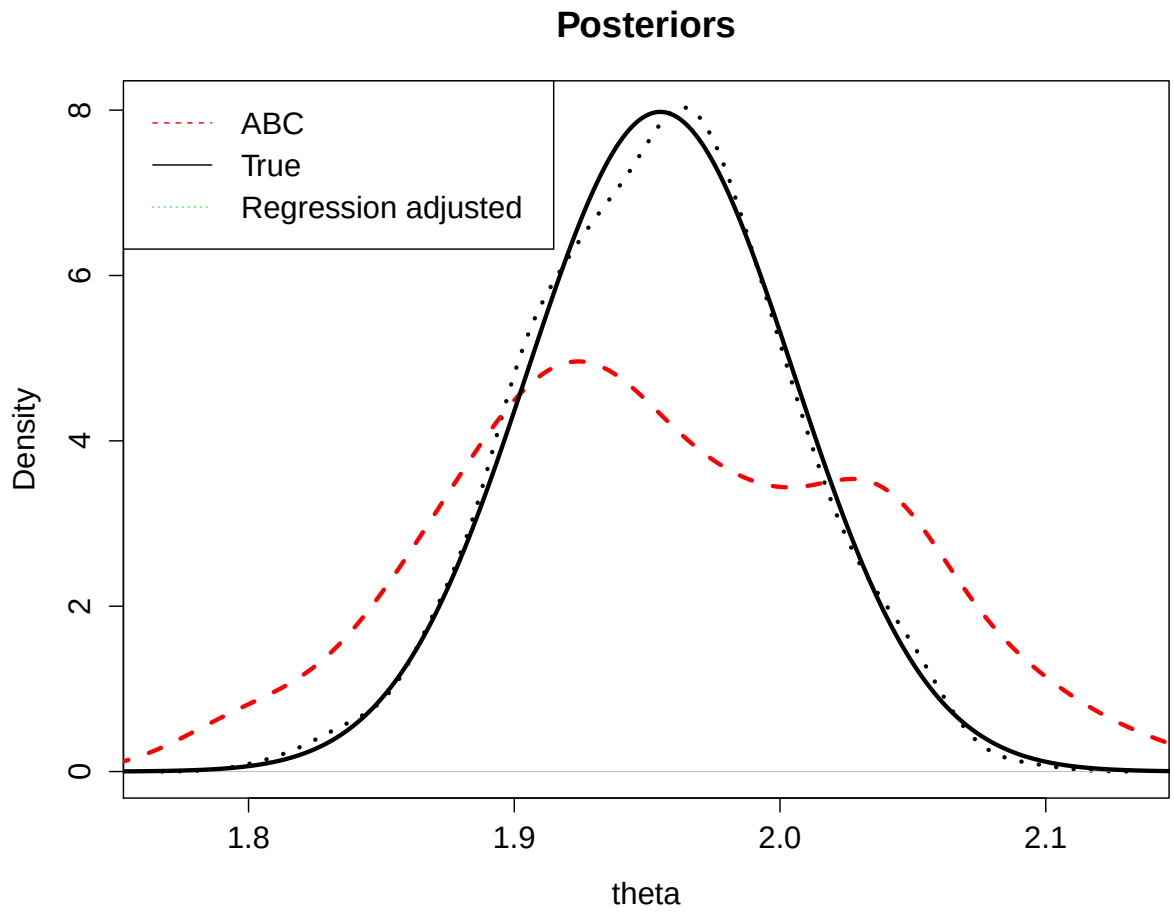
$$\sum (\theta_i - m(s_i))^2 K_\epsilon(s_i - s_{obs})$$

so that observations nearest to s_{obs} are given more weight in the fit.

The empirical residuals are then weighted so that the approximation to the posterior is a weighted particle set

$$\{\theta_i^*, W_i = K_\epsilon(s_i - s_{obs})\}$$
$$\pi(\theta|s_{obs}) = \widehat{m}(s_{obs}) + \sum w_i \delta_{\theta_i^*}(\theta)$$

Normal-normal conjugate model, linear regression



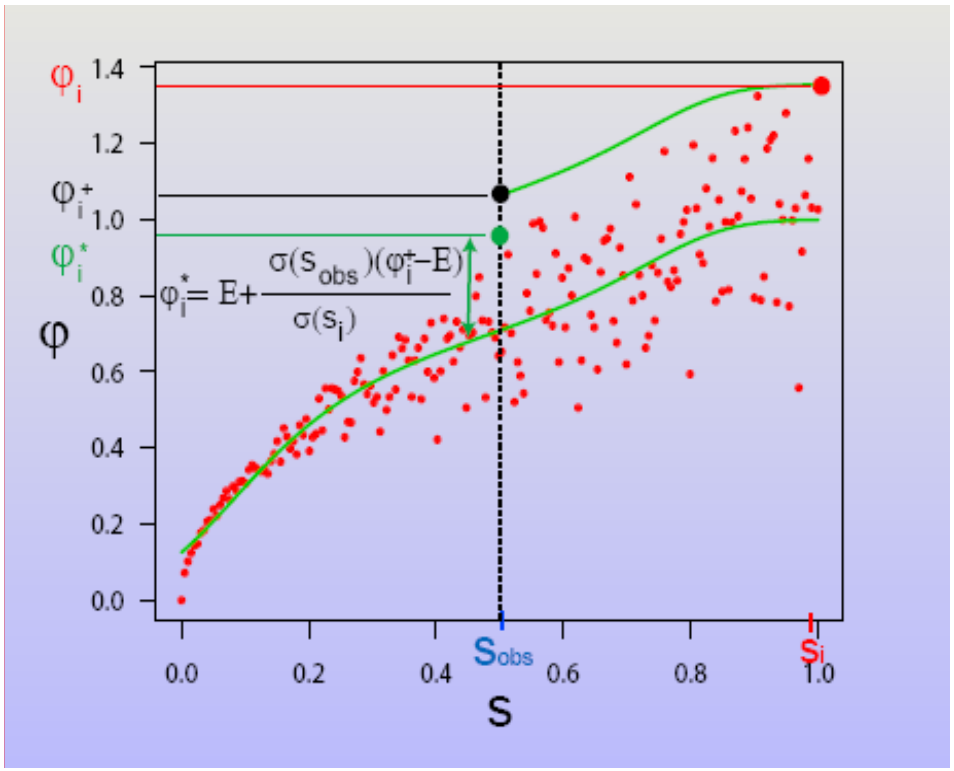
200 data points in both approximations. The regression-adjusted ABC gives a more confident posterior, as the θ_i have been adjusted to account for the discrepancy between s_i and s_{obs}

Extensions: Non-linear models

Blum and Francois 2010 proposed a nonlinear heteroscedastic model

$$\theta_i = m(s_i) + \sigma(s_u)e_i$$

where $m(s) = \mathbb{E}(\theta|s)$ and $\sigma^2(s) = \mathbb{V}\text{ar}(\theta|s)$. They used feed-forward neural networks for both the conditional mean and variance.



$$\theta_i^* = m(s_{obs}) + (\theta_i - \hat{m}(s_i))\,\frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s_i)}$$

Blum 2010 contains estimates of the bias and variance of these estimators: properties of the ABC estimators may seriously deteriorate as $\dim(s)$ increases.

R package `diyABC` implements these methods.

Picture from Michael Blum

Expensive simulators

Motivation

Expensive stochastic simulators exist

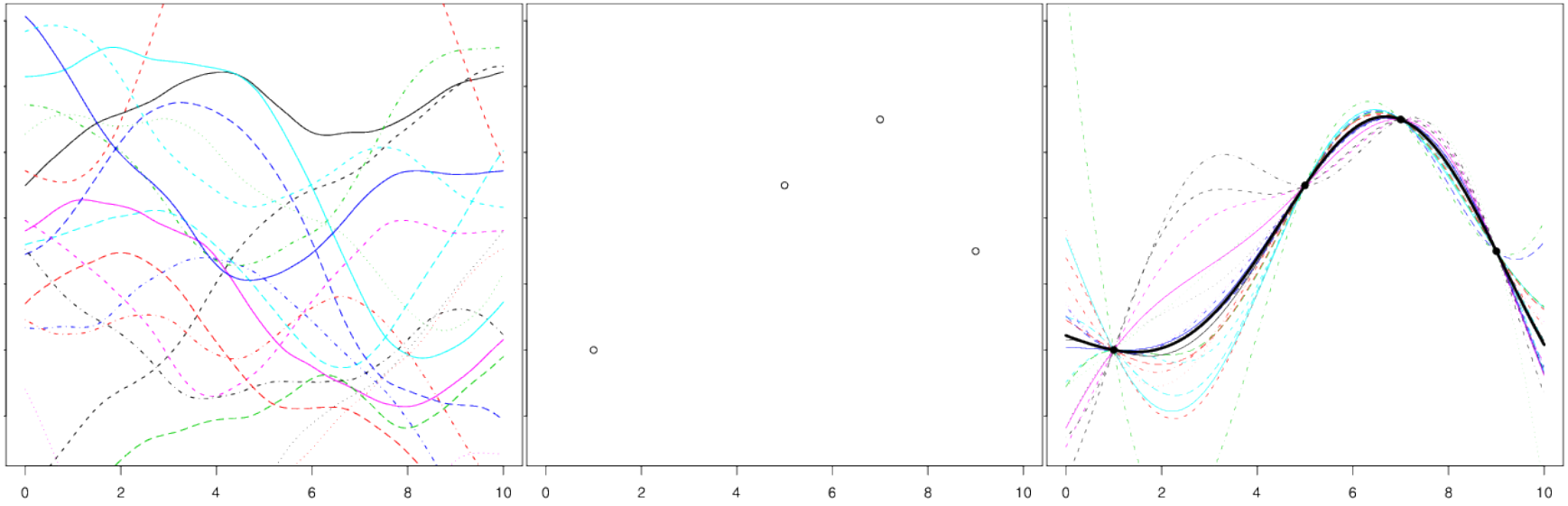
E.g. Cellular Potts model for a human colon crypt

- 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.

If in doubt, use a Gaussian process



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

Emulating likelihood

Wilkinson 2014, Dahlin and Lindsten 2014

Kennedy and O'Hagan built emulators of entire simulator response across all of input space for deterministic functions.

Emulating likelihood

Wilkinson 2014, Dahlin and Lindsten 2014

Kennedy and O'Hagan built emulators of entire simulator response across all of input space for deterministic functions.

If parameter estimation/model selection is the goal, we only need the likelihood function

$$L(\theta) = \pi(D|\theta)$$

which is defined for fixed D.

Instead of modelling the simulator output, we can instead model $L(\theta)$

- A local approximation: D remains fixed, and we only need learn L as a function of θ
- 1d response surface
- But, it can be hard to model.

ABC

One approach is to emulate the synthetic likelihood introduced in Wood 2010.

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

ABC

One approach is to emulate the synthetic likelihood introduced in Wood 2010.

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

- 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*

ABC

One approach is to emulate the synthetic likelihood introduced in Wood 2010.

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

- 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*

An alternative is to emulate the GABC likelihood, or the discrepancy function, or μ_θ and Σ_θ, or ...

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

History matching waves

Wilkinson 2014

The likelihood is too difficult to model, so we model the log-likelihood instead.

$$l(\theta) = \log L(\theta)$$

History matching waves

Wilkinson 2014

The likelihood is too difficult to model, so we model the log-likelihood instead.

$$l(\theta) = \log L(\theta)$$

However, the log-likelihood for a typical problem ranges across too wide a range of values.

Consequently, most GP models will struggle to model the log-likelihood across the parameter space.

History matching waves

Wilkinson 2014

The likelihood is too difficult to model, so we model the log-likelihood instead.

$$l(\theta) = \log L(\theta)$$

However, the log-likelihood for a typical problem ranges across too wide a range of values.

Consequently, most GP models will struggle to model the log-likelihood across the parameter space.

- Introduce waves of history matching, as used in Michael Goldstein's work.
- In each wave, build a GP model that can rule out regions of space as implausible.

Implausibility

Given a model of the likelihood

$$I(\theta) \sim N(m(\theta), \sigma^2)$$

we decide that θ is implausible if

$$m(\theta) + 3\sigma < T$$

Implausibility

Given a model of the likelihood

$$l(\theta) \sim N(m(\theta), \sigma^2)$$

we decide that θ is implausible if

$$m(\theta) + 3\sigma < T$$

- The threshold T can be set in a variety of ways. We use

$$T = \max_{\theta_i} l(\theta_i) - 10$$

for the Ricker model results below,

  - a difference of 10 on the log scale between two likelihoods, means that assigning the θ with the smaller log-likelihood a posterior density of 0 (by saying it is implausible) is a good approximation.

- This still wasn't enough in some problems, so for the first wave we model $\log(-\log \pi(D|\theta))$
- For the next wave, we begin by using the Gaussian processes from the previous waves to decide which parts of the input space are implausible.
- We then extend the design into the not-implaussible range and build a new Gaussian process
- This new GP will lead to a new definition of implausibility
- ...

Example: Ricker Model

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.

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

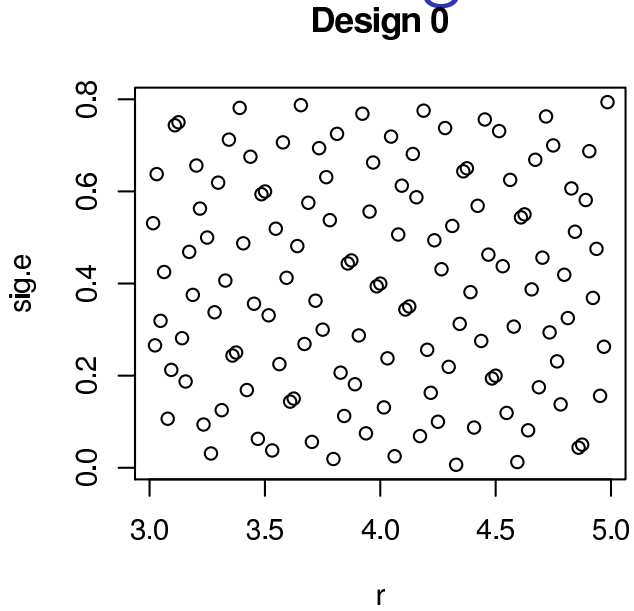
where e_t are independent $N(0, \sigma_e^2)$ process noise

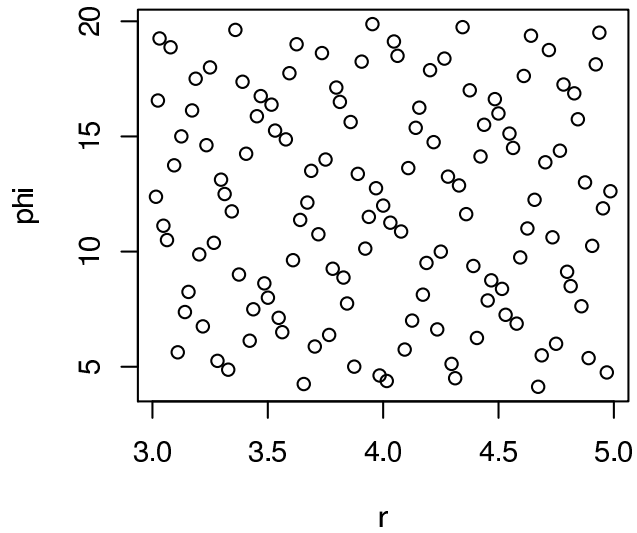
- Assume we observe counts y_t where

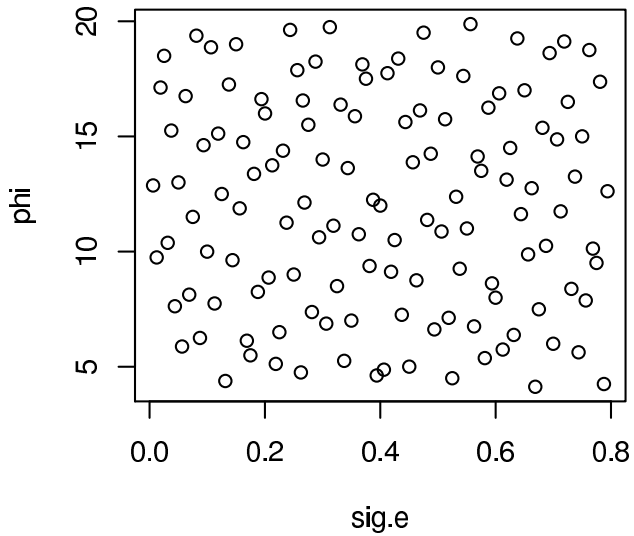
$$y_t \sim Po(\phi N_t)$$

Used in Wood to demonstrate the synthetic likelihood approach.

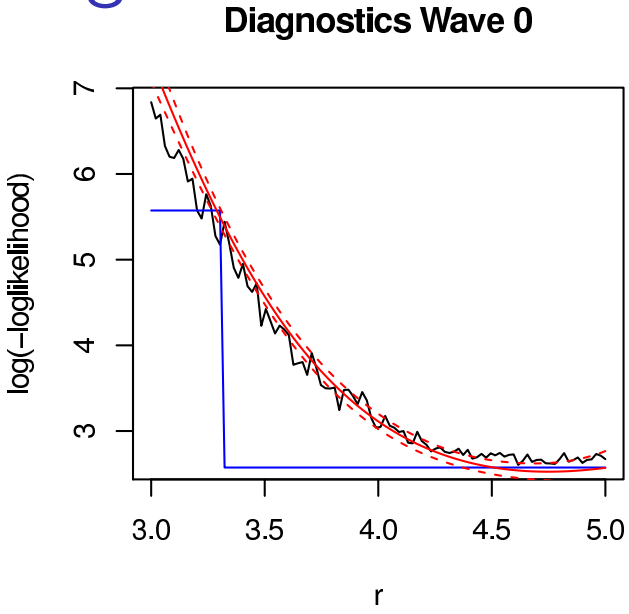
Results - Design 1 - 128 pts

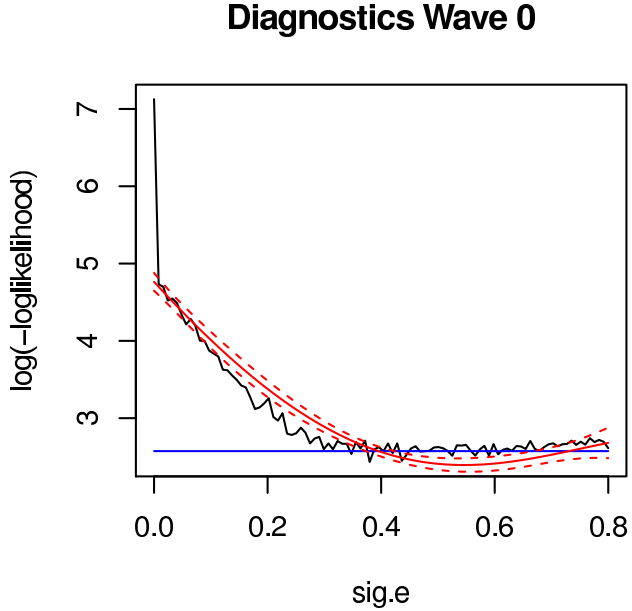


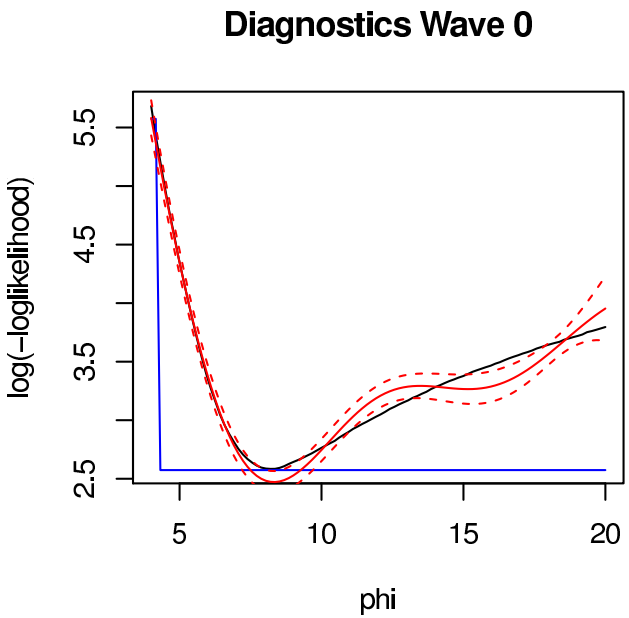




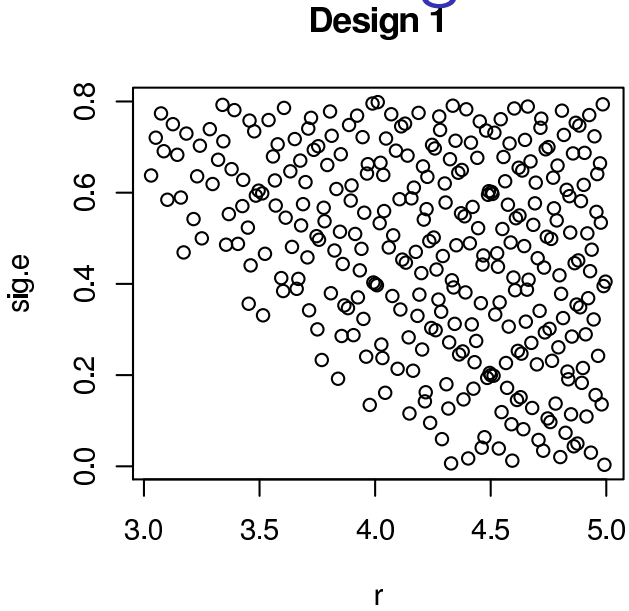
Diagnostics for GP 1 - threshold = 5.6

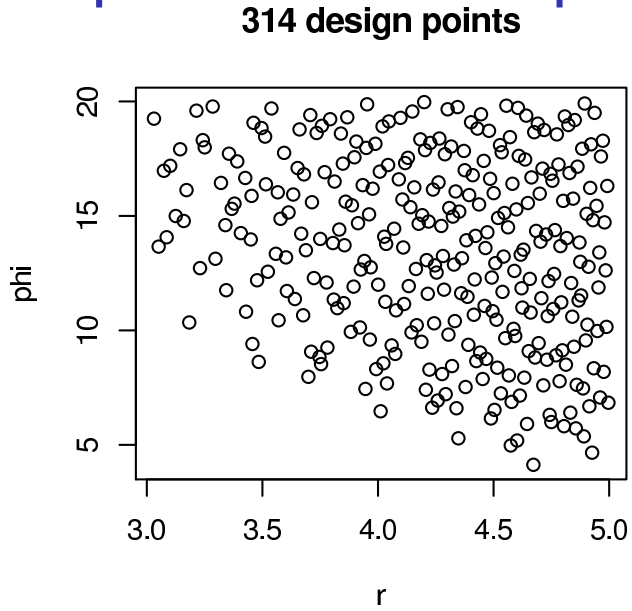


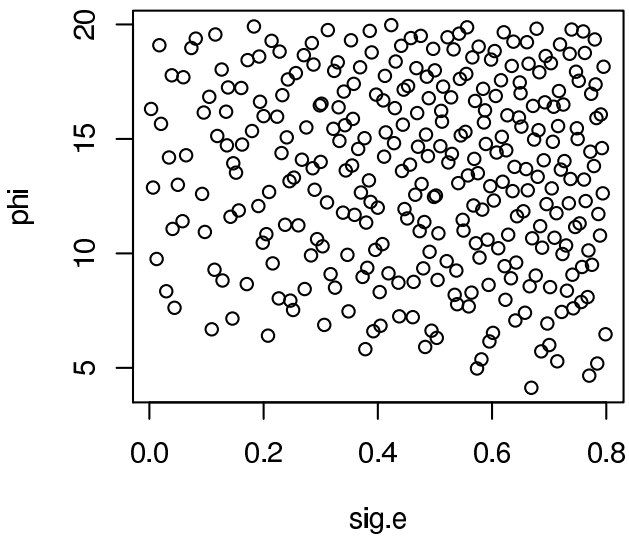




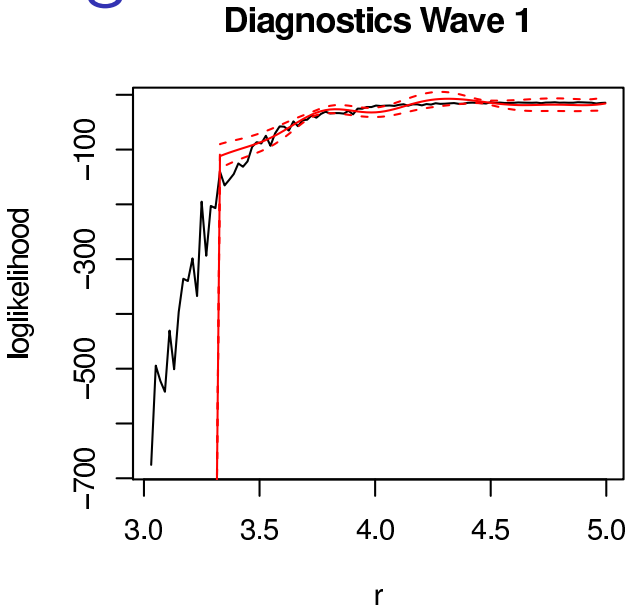
Results - Design 2 - 314 pts - 38% of space implausible

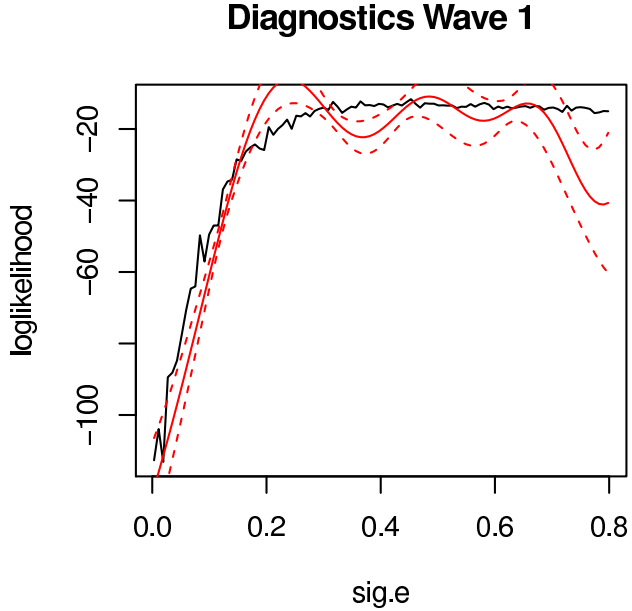


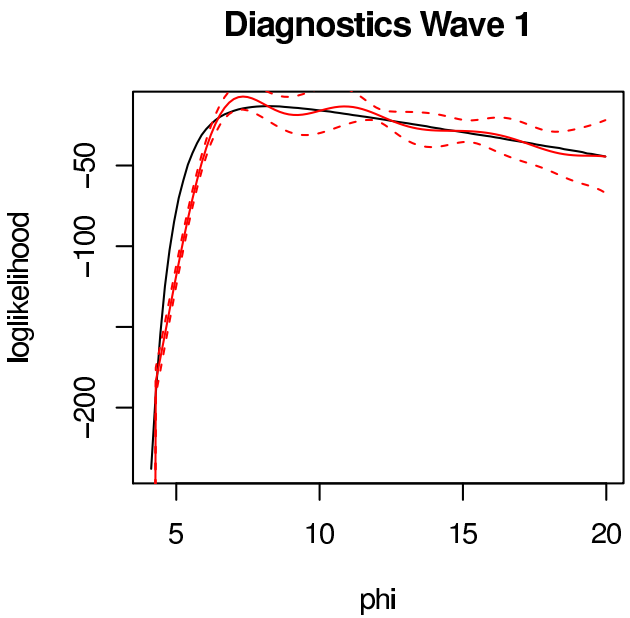




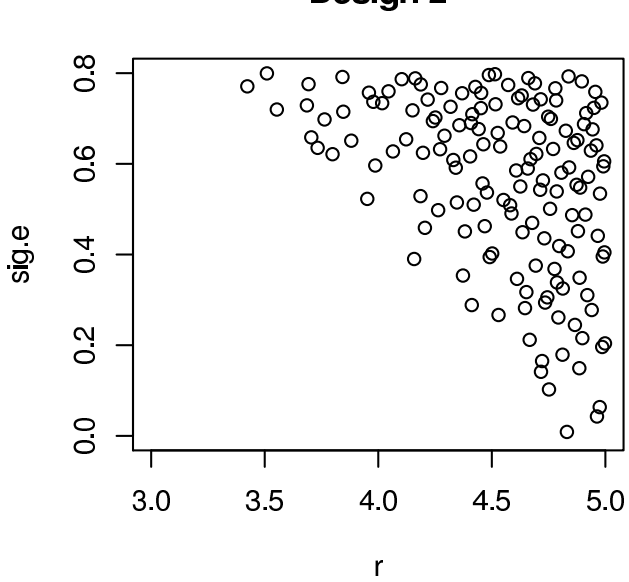
Diagnostics for GP 2 - threshold = -21.8

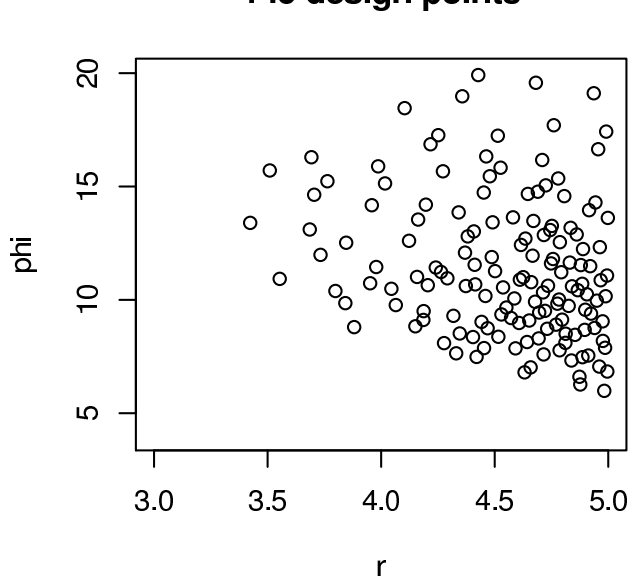


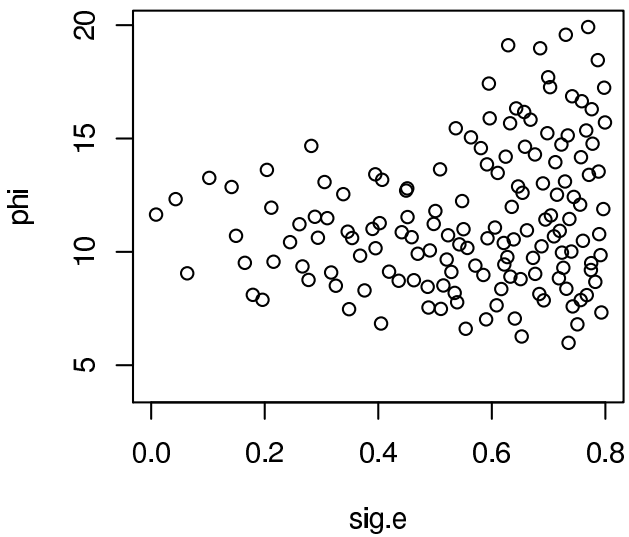




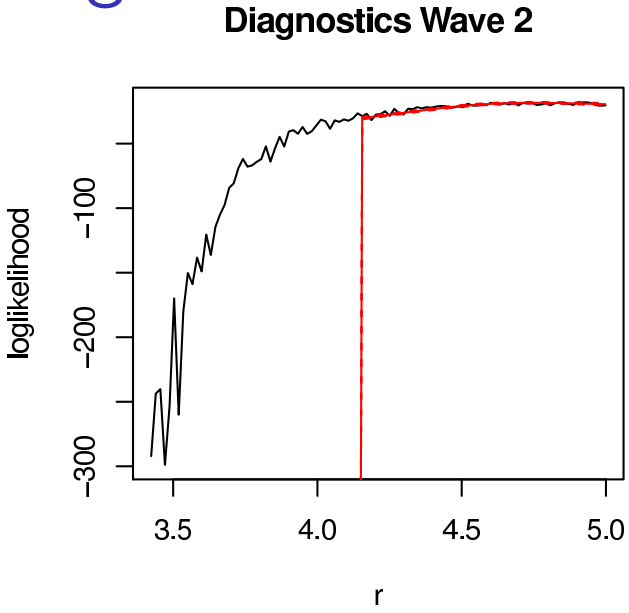
Design 3 - 149 pts - 62% of space implausible

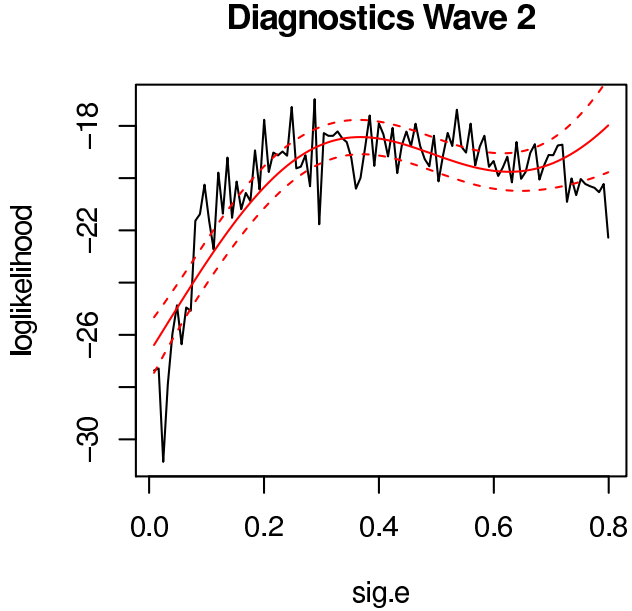


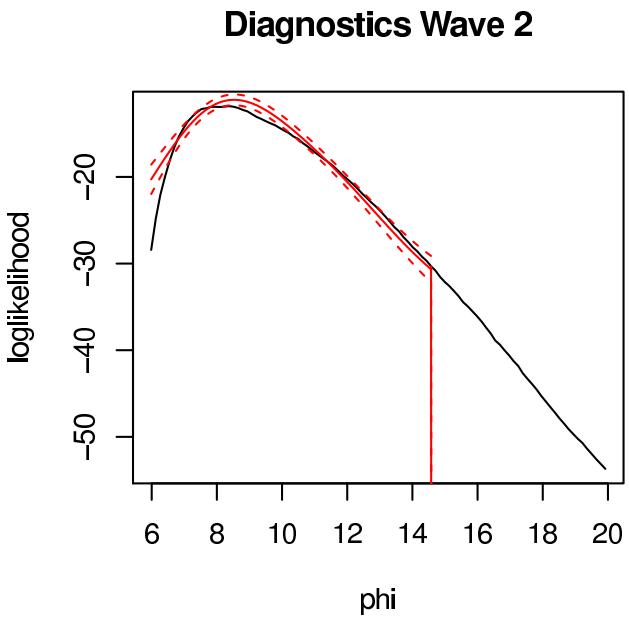




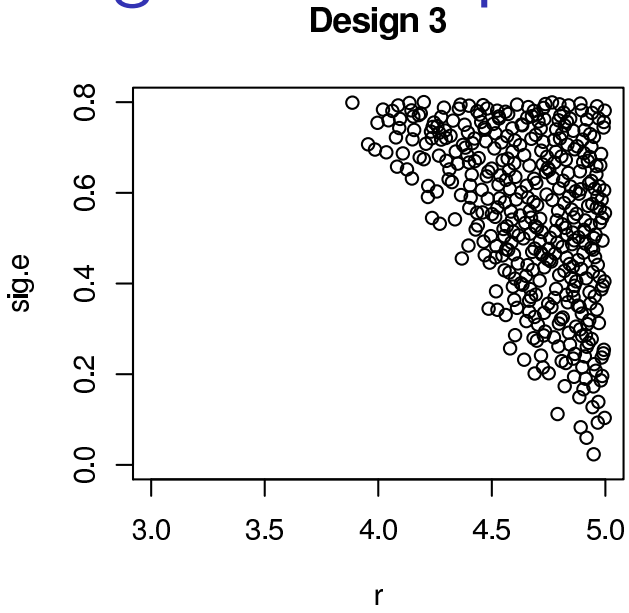
Diagnostics for GP 3 - threshold = -20.7

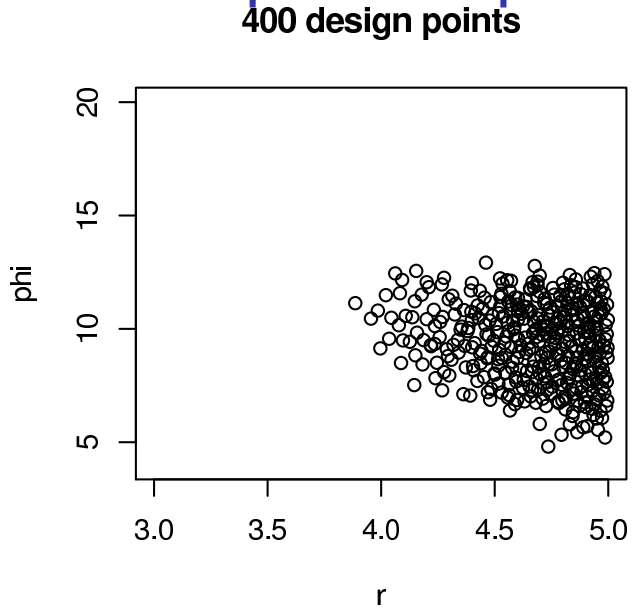


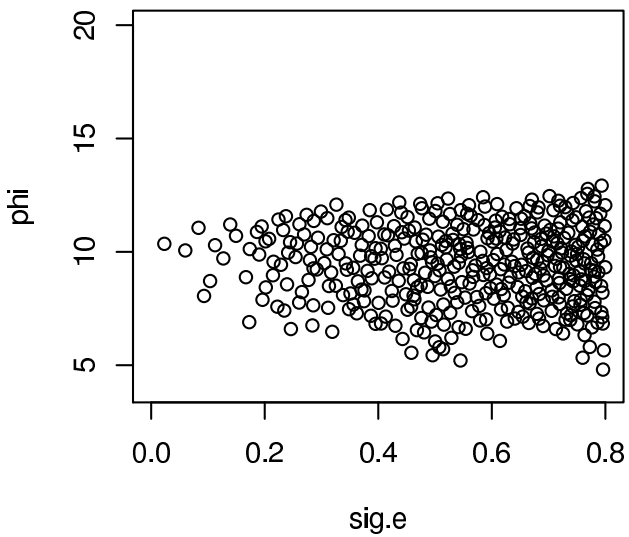




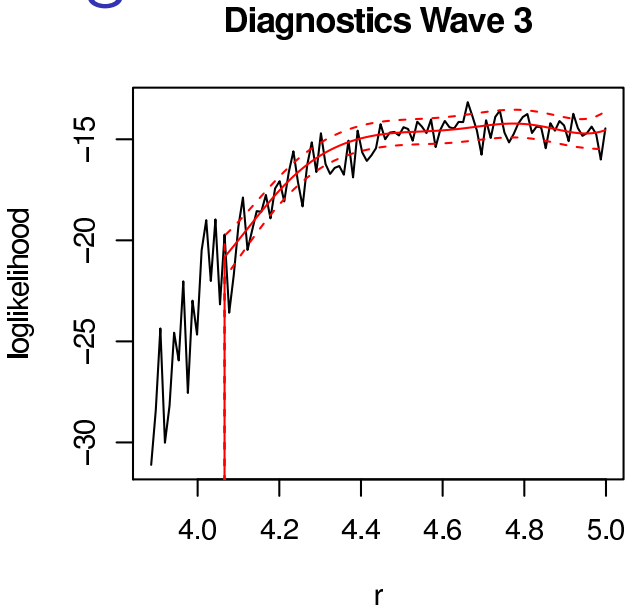
Design 4 - 400 pts - 95% of space implausible

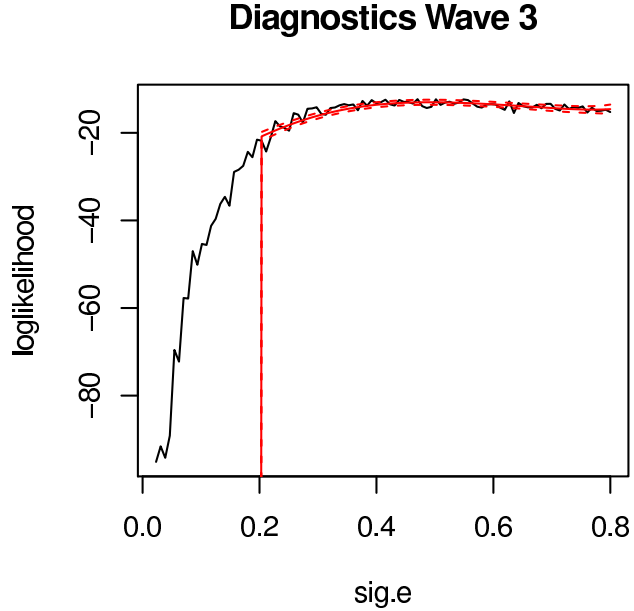


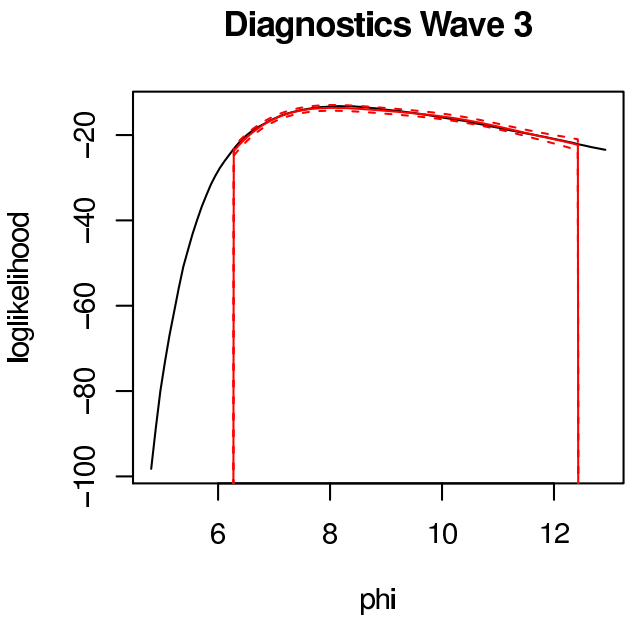




Diagnostics for GP 4 - threshold = -16.4

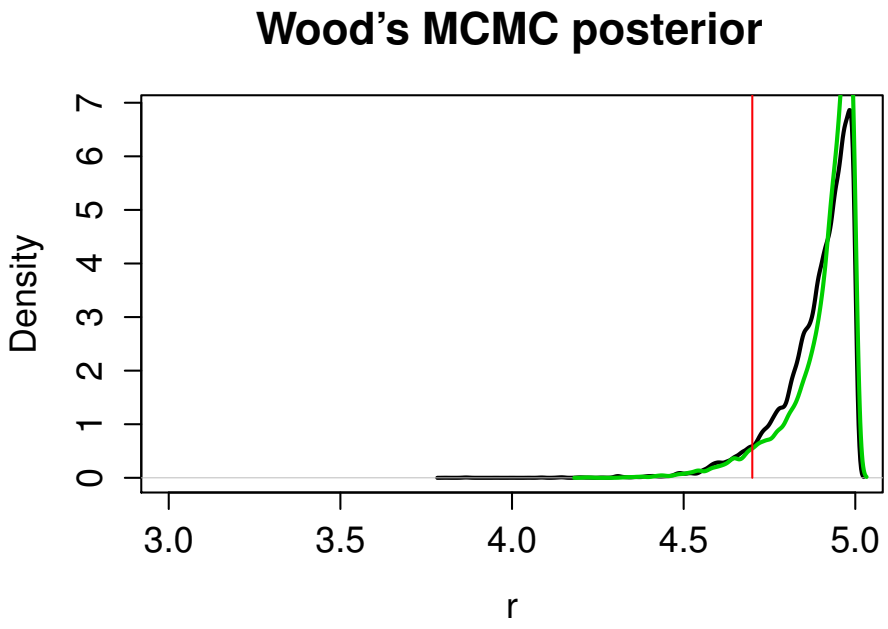


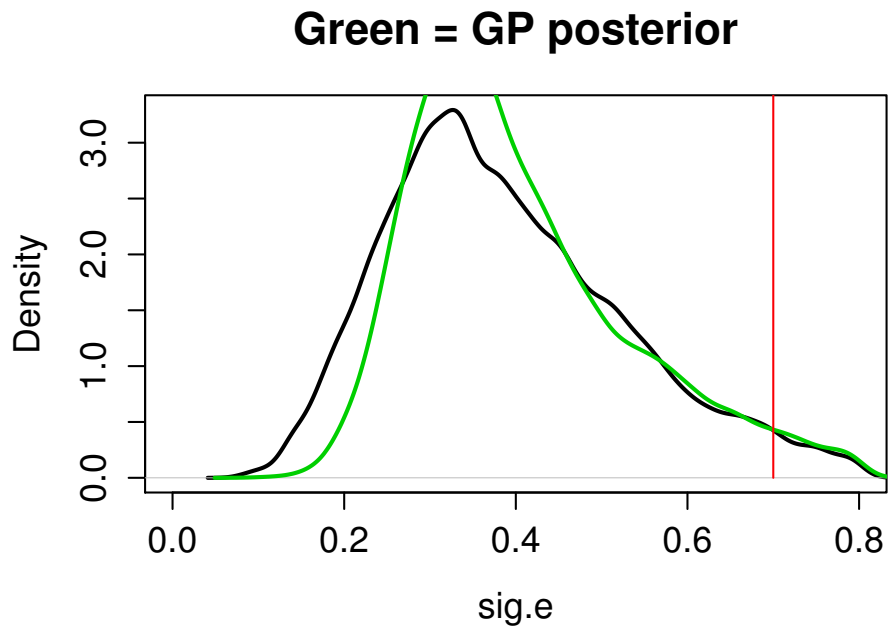


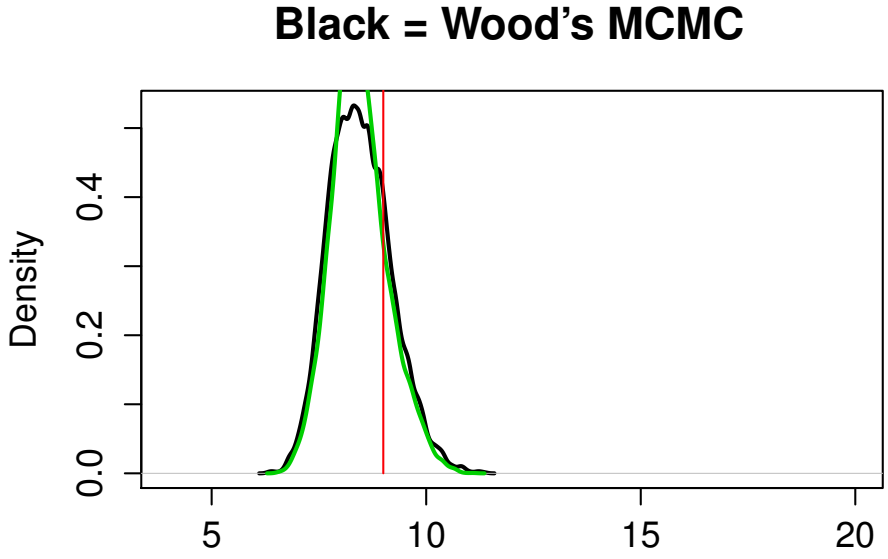


MCMC Results

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.

The ML invasion

Home | About | Resources | Programs | Live Stream | Videos | Services | Publications | Search | Contact

- Objectives
- Confirmed Participants
- Press Release
- Meeting Facilities
- Mailing List

Validating and Expanding Approximate Bayesian Computation Methods (17w5025)

Arriving in Banff, Alberta Sunday, February 19 and departing Friday February 24, 2017

Organizers

Christian Robert (Université Paris-Dauphine)

Luke Bornn (Simon Fraser University)

Gael Martin (Monash University)

Jukka Corander (University of Helskinki)

Dennis Prangle (University of Newcastle)

Richard Wilkinson (University of Sheffield)

Conclusions

ABC allows inference in models for which it would otherwise be impossible.

- not a silver bullet - if likelihood methods possible, use them instead.

Algorithms and post-hoc regression can greatly improve computational efficiency, but computation is still usually the limiting factor. Challenge is

- to develop more efficient methods to allow inference in more expensive models
- find better ways to more efficiently summarize the data

Conclusions

ABC allows inference in models for which it would otherwise be impossible.

- not a silver bullet - if likelihood methods possible, use them instead.

Algorithms and post-hoc regression can greatly improve computational efficiency, but computation is still usually the limiting factor. Challenge is

- to develop more efficient methods to allow inference in more expensive models
- find better ways to more efficiently summarize the data

Conclusions

ABC allows inference in models for which it would otherwise be impossible.

- not a silver bullet - if likelihood methods possible, use them instead.

Algorithms and post-hoc regression can greatly improve computational efficiency, but computation is still usually the limiting factor. Challenge is

- to develop more efficient methods to allow inference in more expensive models
- find better ways to more efficiently summarize the data

Thank you for listening!

References - basics

Included in order of appearance in tutorial, rather than importance! Far from exhaustive - apologies to those I've missed

- Murray, Ghahramani, MacKay, *NIPS*, 2012
- Tanaka, Francis, Luciani and Sisson, *Genetics* 2006.
- Wilkinson, Tavare, *Theoretical Population Biology*, 2009,
- Neal and Huang, *arXiv*, 2013.
- Beaumont, Zhang, Balding, *Genetics* 2002
- Tavare, Balding, Griffiths, *Genetics* 1997
- Diggle, Gratton, *JRSS Ser. B*, 1984
- Rubin, *Annals of Statistics*, 1984
- Wilkinson, *SAGMB* 2013.
- Fearnhead and Prangle, *JRSS Ser. B*, 2012
- Kennedy and O'Hagan, *JRSS Ser. B*, 2001

References - algorithms

- Marjoram, Molitor, Plagnol, Tavarè, *PNAS*, 2003
- Sisson, Fan, Tanaka, *PNAS*, 2007
- Beaumont, Cornuet, Marin, Robert, *Biometrika*, 2008
- Toni, Welch, Strelkowa, Ipsen, Stumpf, *Interface*, 2009.
- Del Moral, Doucet, *Stat. Comput.* 2011
- Drovandi, Pettitt, *Biometrics*, 2011.
- Lee, *Proc 2012 Winter Simulation Conference*, 2012.
- Lee, Latuszynski, *arXiv*, 2013.
- Del Moral, Doucet, Jasra, *JRSS Ser. B*, 2006.
- Sisson and Fan, *Handbook of MCMC*, 2011.

References - links to other algorithms

- Craig, Goldstein, Rougier, Seheult, *JASA*, 2001
- Fearnhead and Prangle, *JRSS Ser. B*, 2011.
- Wood *Nature*, 2010
- Nott and Marshall, *Water resources research*, 2012
- Nott, Fan, Marshall and Sisson, *arXiv*, 2012.

GP-ABC:

- Wilkinson, *arXiv*, 2013
- Meeds and Welling, *arXiv*, 2013.

References - regression adjustment

- Beaumont, Zhang, Balding, *Genetics*, 2002
- Blum, Francois, *Stat. Comput.* 2010
- Blum, *JASA*, 2010
- Leuenberger, Wegmann, *Genetics*, 2010

References - summary statistics

- Blum, Nunes, Prangle, Sisson, *Stat. Sci.*, 2012
- Joyce and Marjoram, *Stat. Appl. Genet. Mol. Biol.*, 2008
- Nunes and Balding, *Stat. Appl. Genet. Mol. Biol.*, 2010
- Fearnhead and Prangle, *JRSS Ser. B*, 2011
- Wilkinson, PhD thesis, University of Cambridge, 2007
- Grelaud, Robert, Marin *Comptes Rendus Mathematique*, 2009
- Robert, Cornuet, Marin, Pillai *PNAS*, 2011
- Didelot, Everitt, Johansen, Lawson, *Bayesian analysis*, 2011.