Approximate Bayesian computation (ABC) and the challenge of big simulation

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

September 3, 2014

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

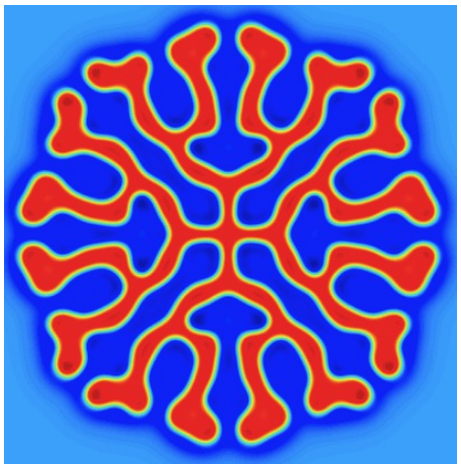
- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- The inverse-problem: observe data D, estimate parameter values θ which explain the data.

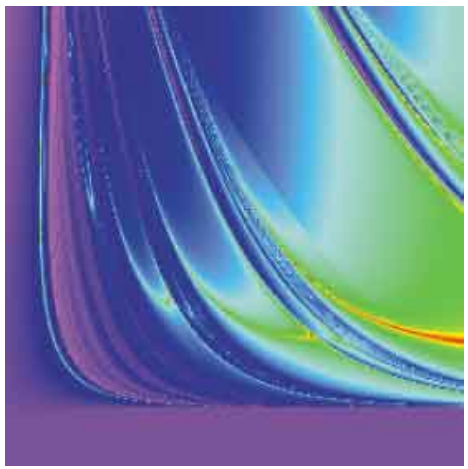
The Bayesian approach is to find the posterior distribution

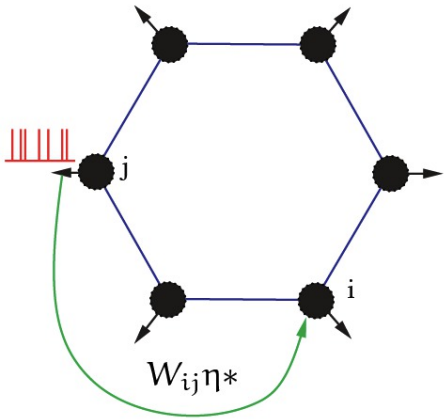
$$\pi(\theta|D) \propto \pi(\theta)\pi(D|\theta)$$

posterior $\propto$

prior $\times$ 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. Summary statistics
vi. Accelerating ABC using meta-models

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:

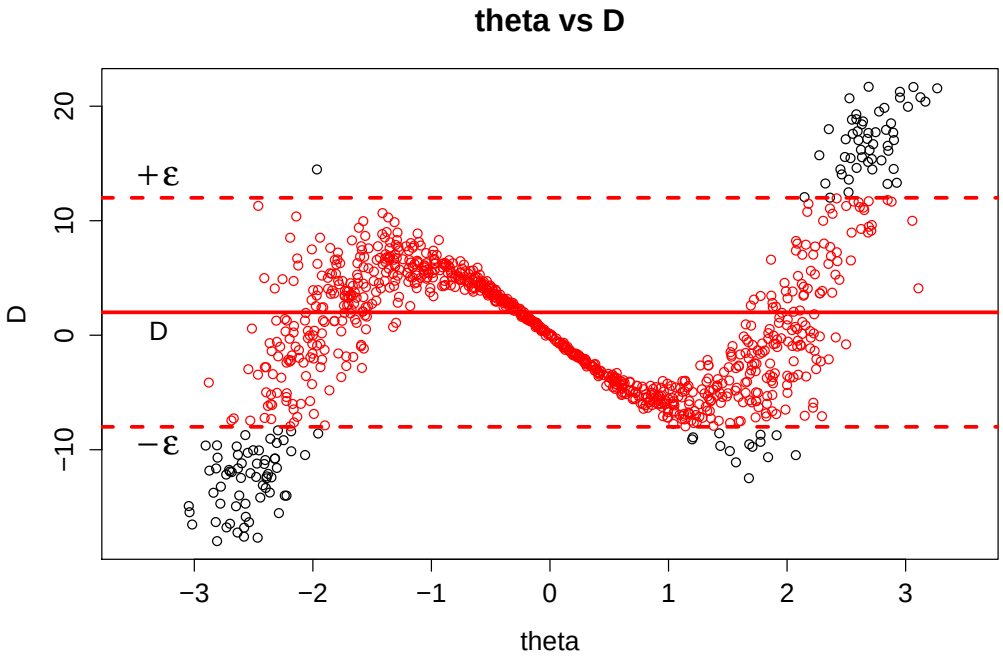
Uniform Rejection Algorithm

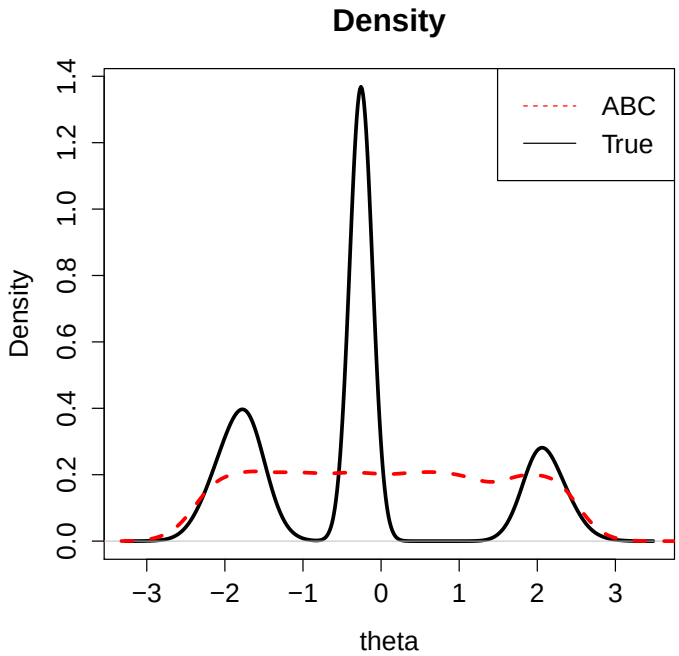
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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

$\epsilon = 10$

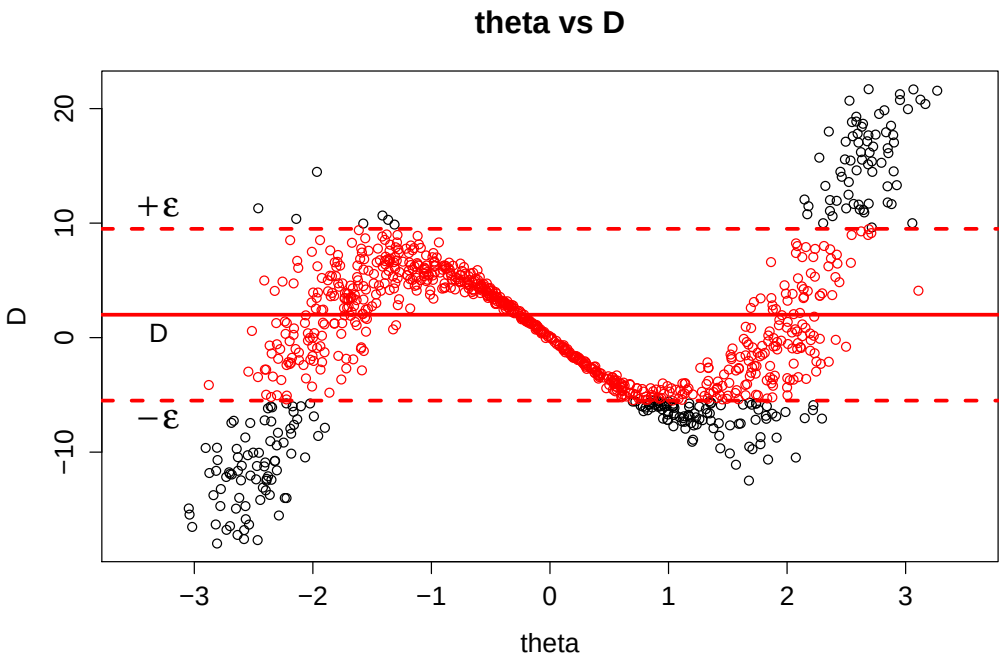


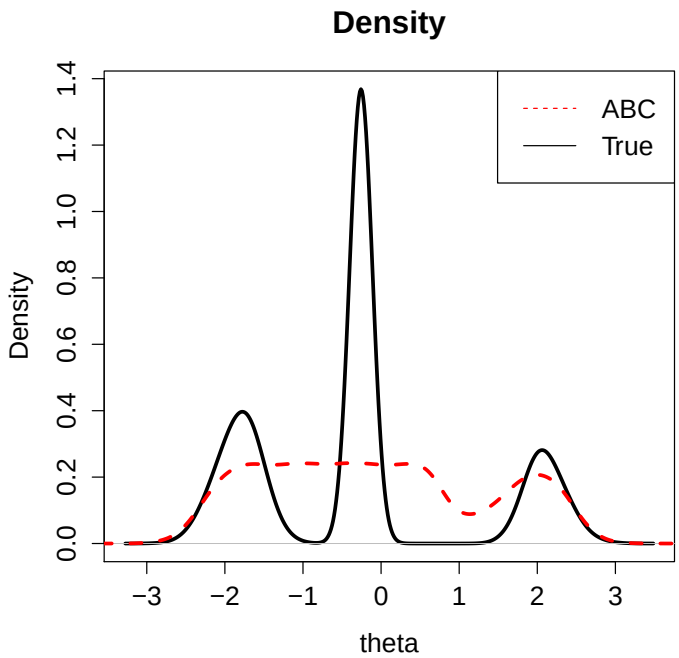


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

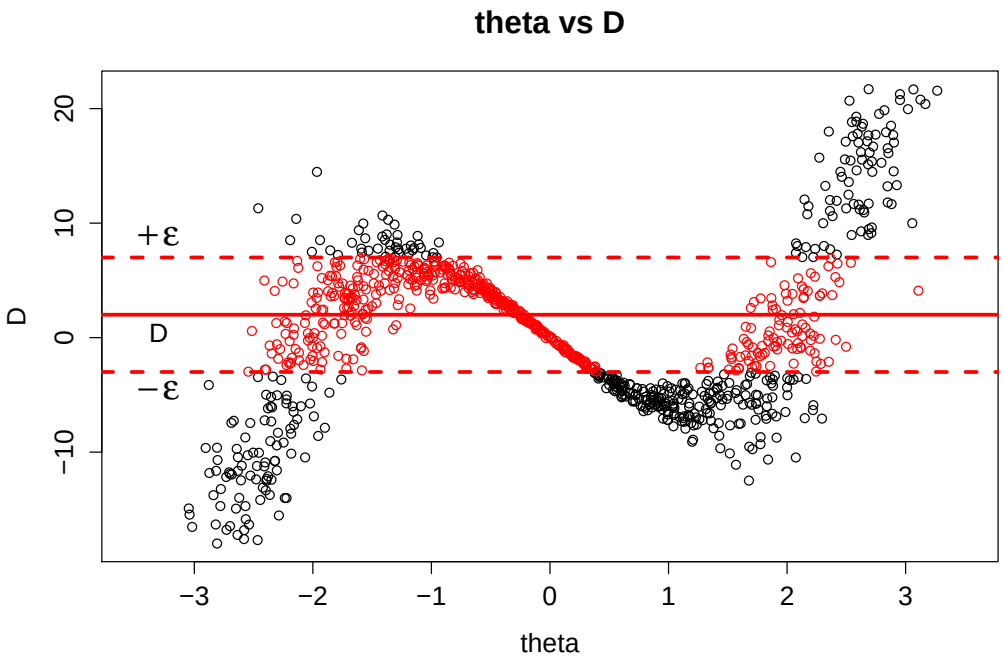
$$\rho(D, X) = |D - X|, \qquad D = 2$$

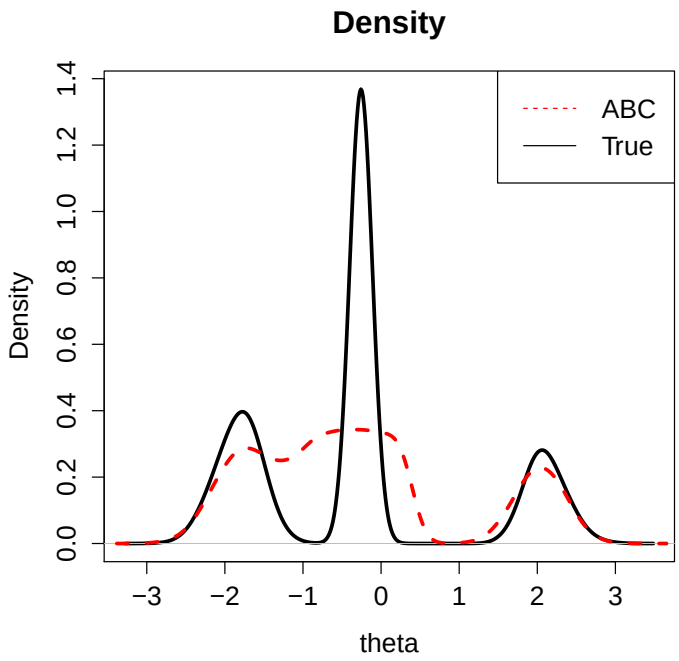
$\epsilon = 7.5$


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

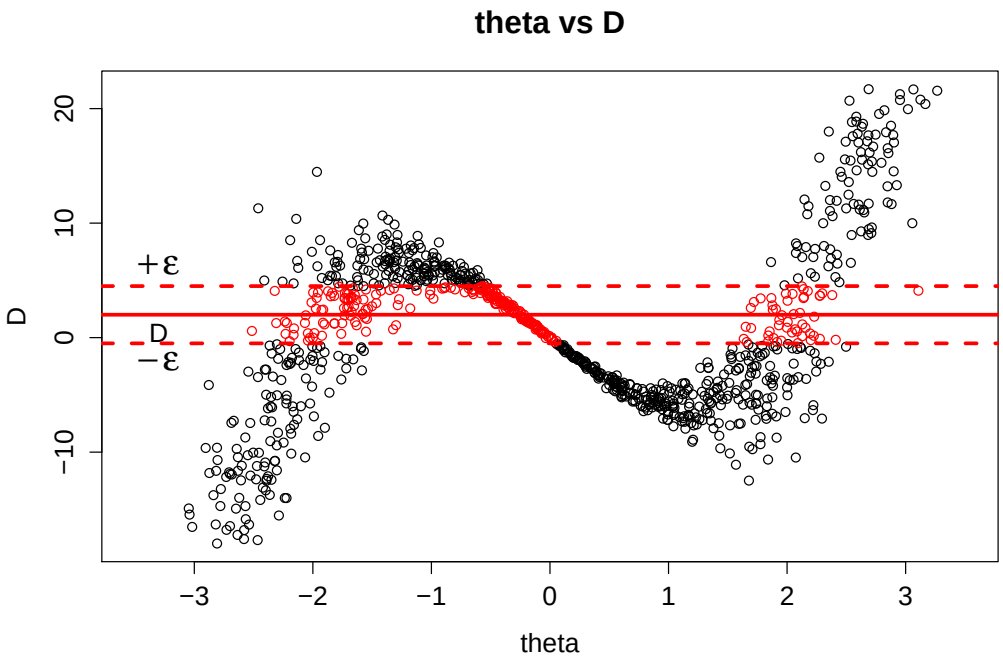

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

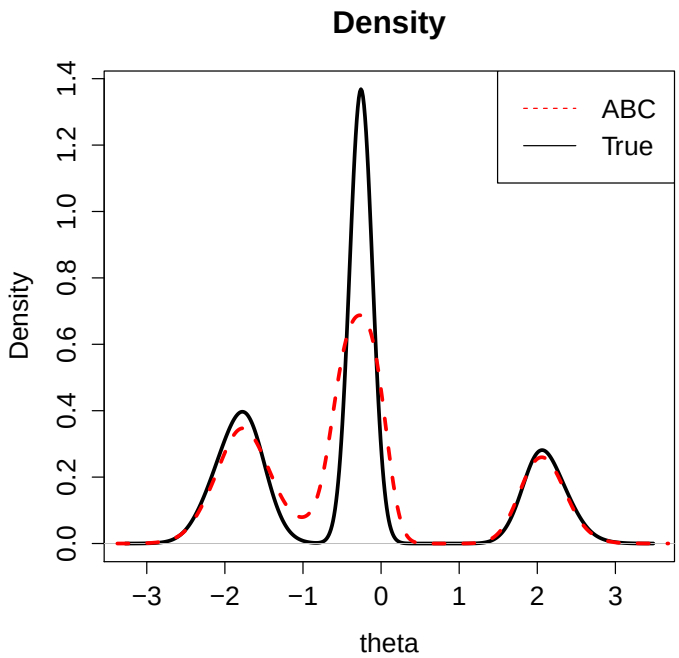
$\epsilon = 5$


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

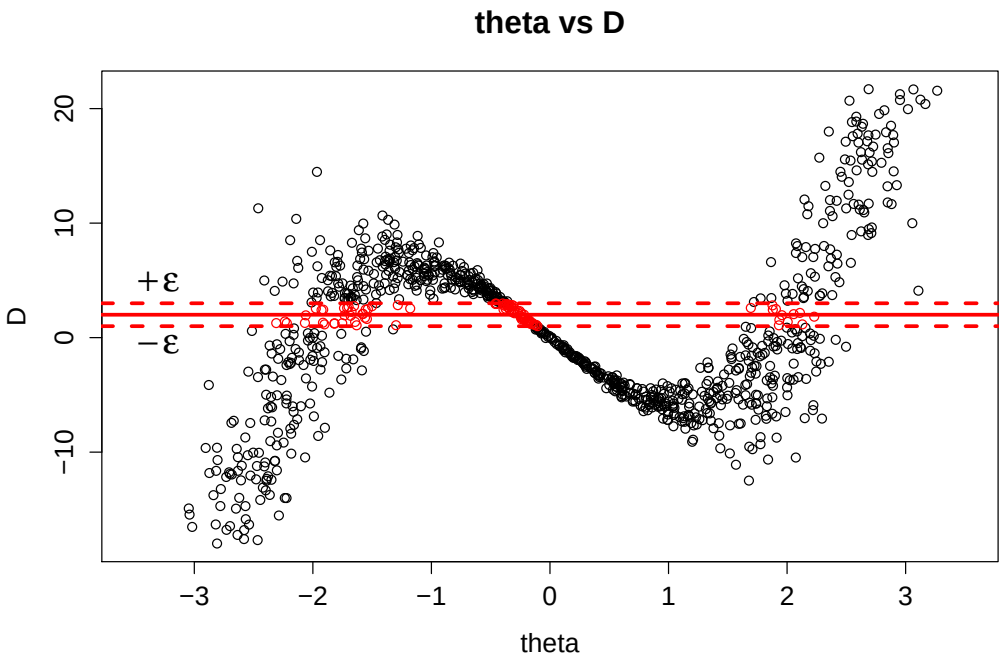

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

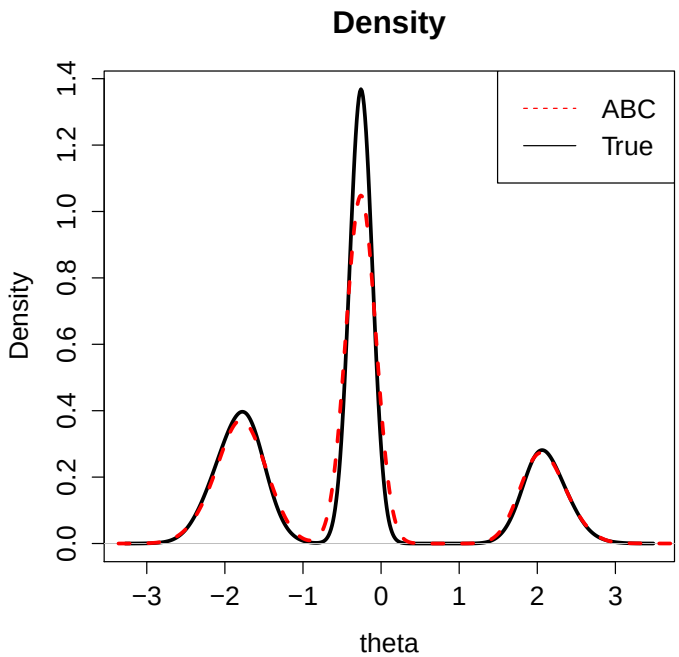
$\epsilon = 2.5$


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

What is ABC doing?

We can think about ABC in two ways:

- Algorithmically:

- Probabilistically:

What is ABC doing?

We can think about ABC in two ways:

- Algorithmically:
  - find a good metric, tolerance and summary etc, to minimise the approximation error
- Probabilistically:
  - Given algorithmic choices, what model does ABC correspond to?, and how should this inform our choices?

ABC as a probability model

Wilkinson 2008

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

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 $X \sim \pi(x|\theta)$ (ie $(\theta, X) \sim g(\cdot)$)
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 $X \sim \pi(x|\theta)$ (ie $(\theta, X) \sim g(\cdot)$)
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 (see Neal *et al.* 2014 for an alternative).

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 (see Neal *et al.* 2014 for an alternative).

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 (see Neal *et al.* 2014 for an alternative).

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.

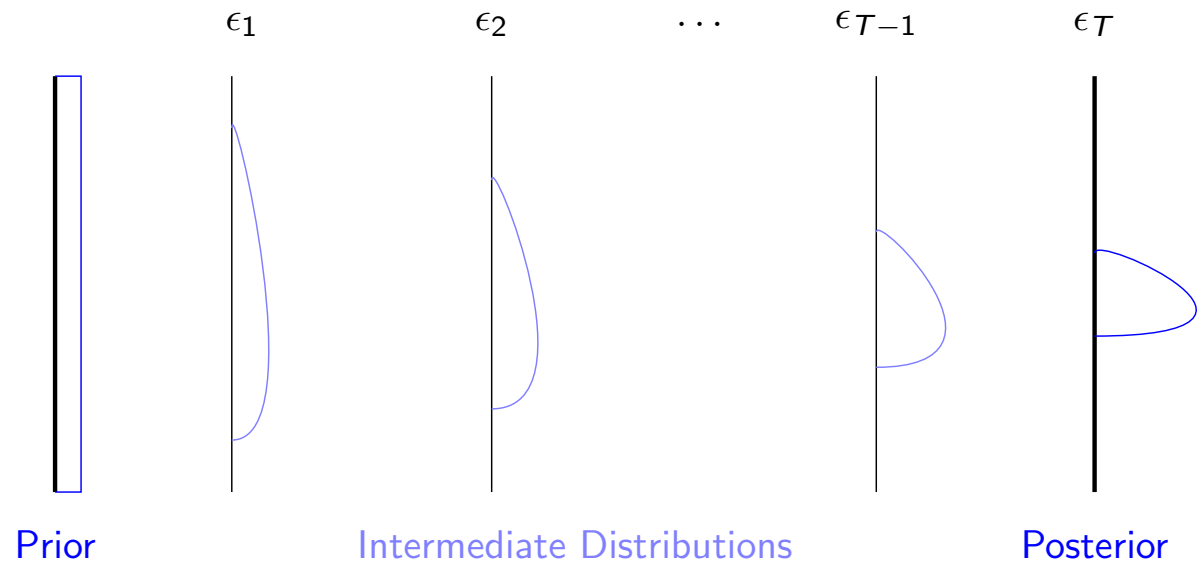
We aim to sample N particles successively from a sequence of distributions

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

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

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

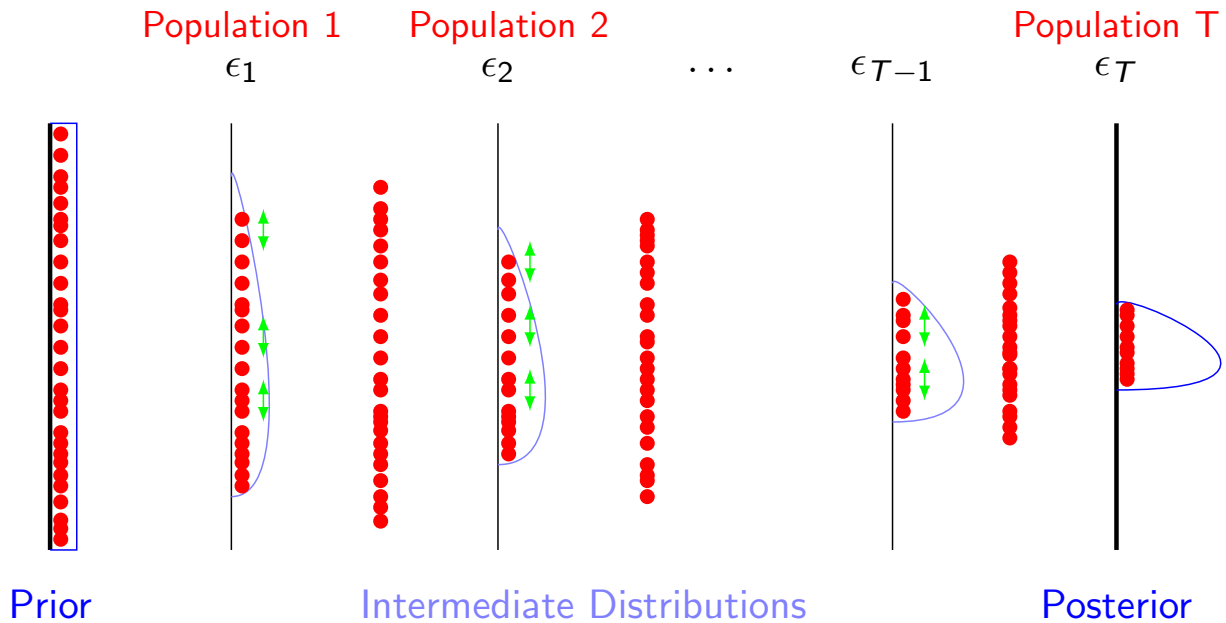


Picture from Toni and Stumpf 2010 tutorial

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

Links to other approaches

History-matching

Craig *et al.* 2001, Vernon *et al.* 2010

ABC can be seen as a probabilistic version of history matching. History matching is used in the analysis of computer experiments to rule out regions of space as implausible.

1. Relate the simulator to the system

$$\zeta = f(\theta) + \epsilon$$

   where ϵ is our simulator discrepancy

2. Relate the system to the data (e represents measurement error)

$$D = \zeta + e$$

3. Declare θ implausible if, e.g.,

$$\| D - \mathbb{E}f(\theta) \| > 3\sigma$$

   where σ^2 is the combined variance implied by the emulator, discrepancy and measurement error.

History-matching

If θ is not implausible we don't discard it. The result is a region of space that we can't rule out at this stage of the history-match.

Usual to go through several stages of history matching.

- History matching can be seen as a principled version of ABC - lots of thought goes into the link between simulator and reality.
- The result of history-matching may be that there is no not-implausible region of parameter space
  - Go away and think harder - something is misspecified
  - This can also happen in rejection ABC.
  - In contrast, MCMC will always give an answer, even if the model is terrible.
- The method is non-probabilistic - it just gives a set of not-implausible parameter values. Probabilistic calibration can be done subsequently.

Other algorithms

- The synthetic likelihood approach of Wood 2010 is an ABC algorithm, but using sample mean μ_θ and covariance Σ_θ of the summary of $f(\theta)$ run n times at θ, and assuming

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

- (Generalized Likelihood Uncertainty Estimation) GLUE approach of Keith Beven in hydrology can also be interpreted as an ABC algorithm - see Nott and Marshall 2012

Regression Adjustment

References:

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

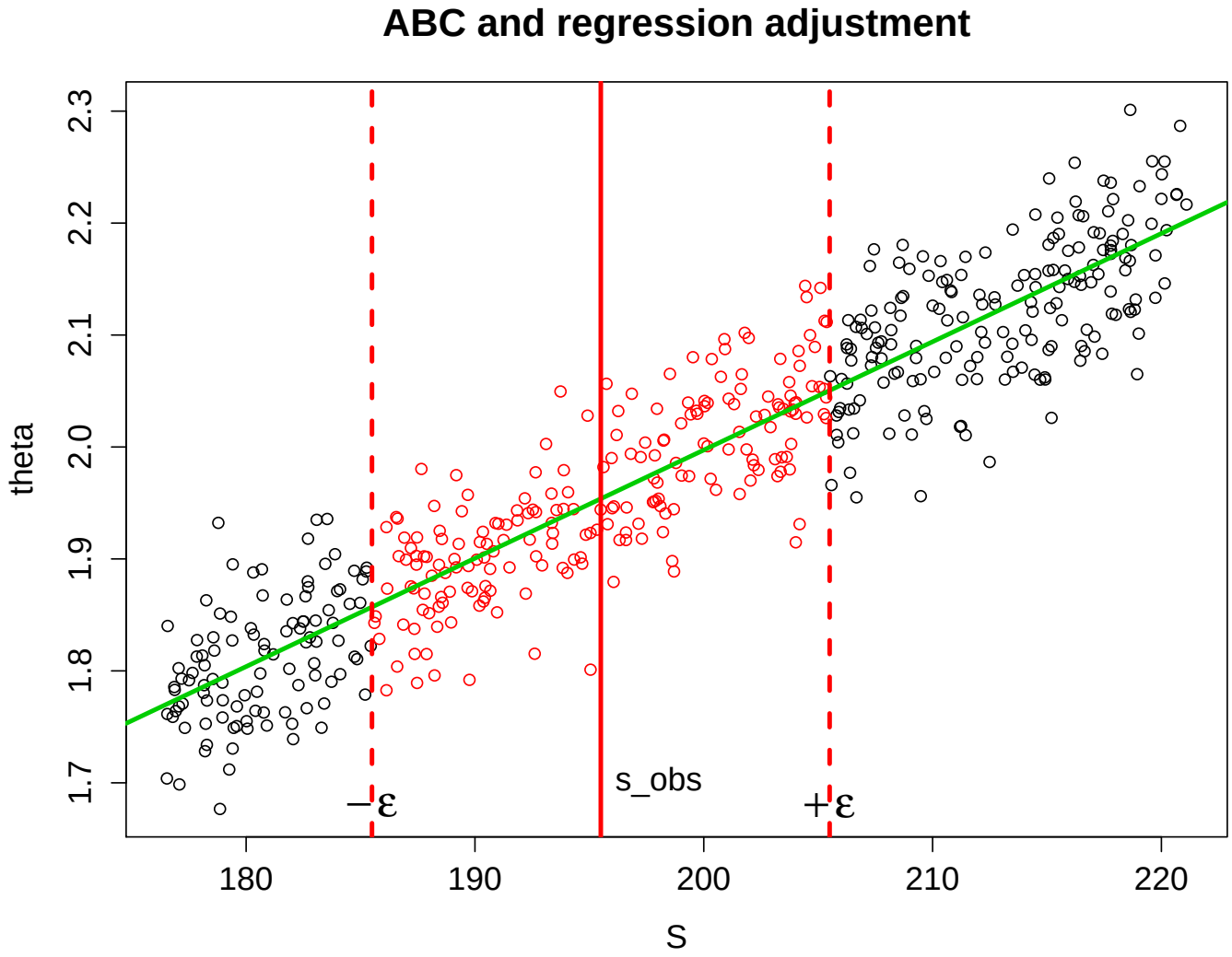
Regression Adjustment

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}.

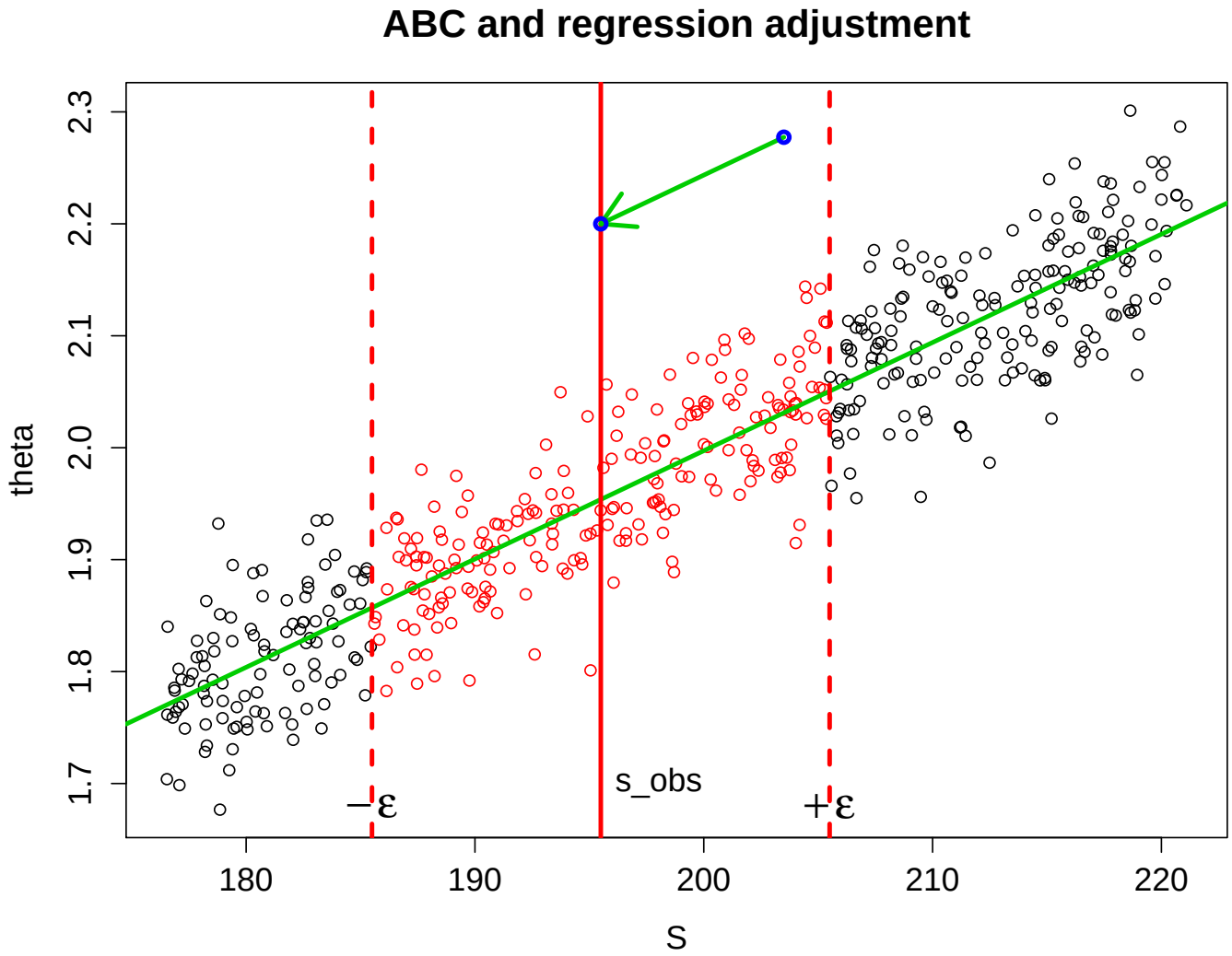
Two key ideas

- 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

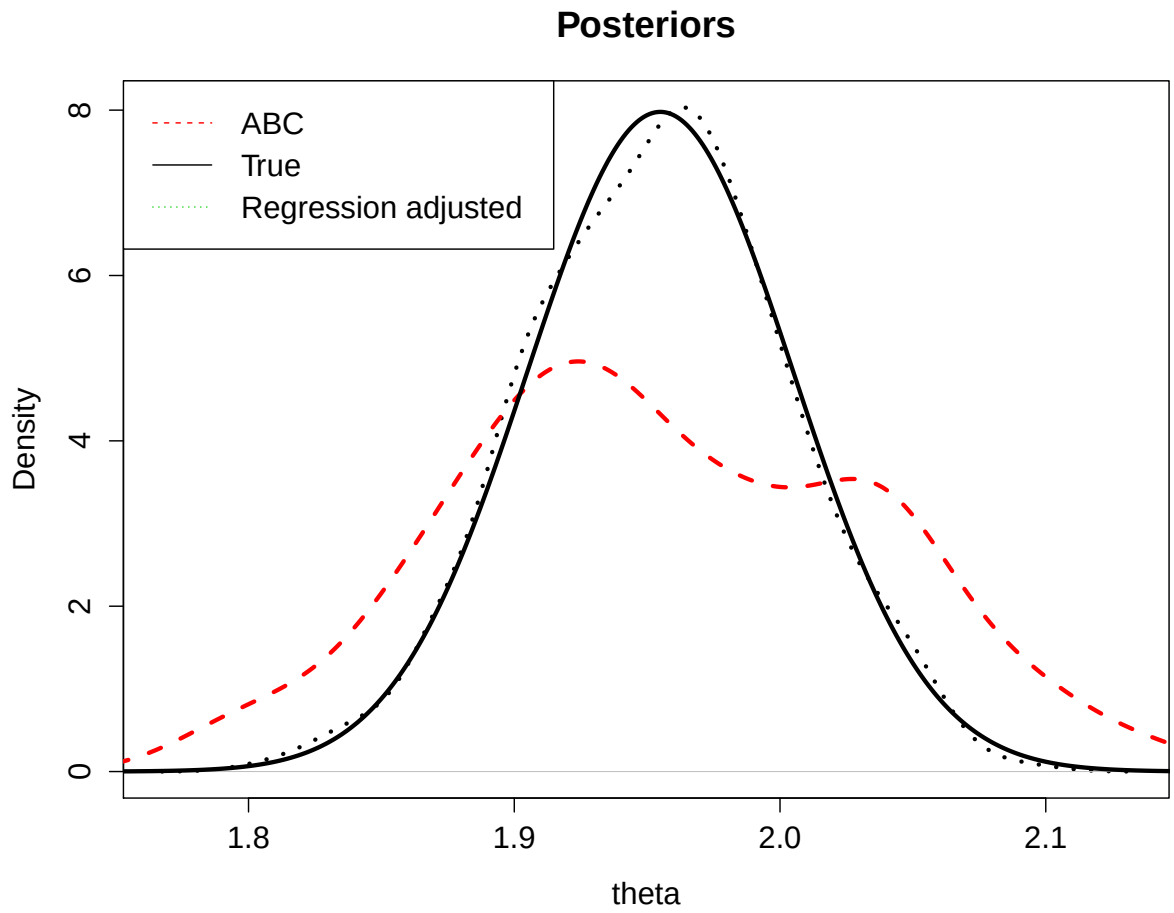
$$\sum (\theta_i - m(s_i))^2 K_\epsilon(s_i - s_{obs})$$

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



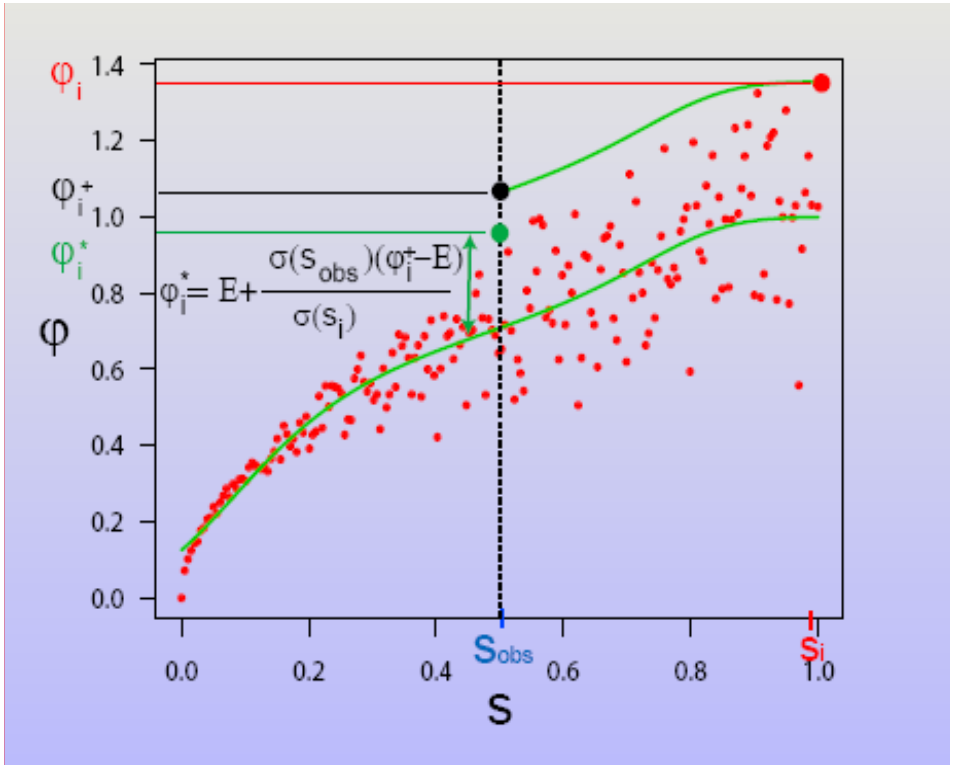
200 data points in both approximations. The regression-adjusted ABC gives a more confident posterior, as the θ_i have been adjusted to account for the discrepancy between s_i and s_{obs}

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

Summary Statistics

References:

- Blum, Nunes, Prangle and Sisson 2012
- Joyce and Marjoram 2008
- Nunes and Balding 2010
- Fearnhead and Prangle 2012
- Robert *et al.* 2011

Error trade-off

Blum, Nunes, Prangle, Fearnhead 2012

The error in the ABC approximation can be broken into two parts

1. Choice of summary:

$$\pi(\theta|D) \stackrel{?}{\approx} \pi(\theta|S(D))$$

Error trade-off

Blum, Nunes, Prangle, Fearnhead 2012

The error in the ABC approximation can be broken into two parts

1. Choice of summary:

$$\pi(\theta|D) \stackrel{?}{\approx} \pi(\theta|S(D))$$

2. Use of ABC acceptance kernel:

$$\pi(\theta|s_{obs}) \stackrel{?}{\approx} \pi_{ABC}(\theta|s_{obs}) = \int \pi(\theta, s|s_{obs})\mathrm{d}s$$

$$\propto \int \pi_{\epsilon}(s_{obs}|S(x))\pi(x|\theta)\pi(\theta)\mathrm{d}x$$

Error trade-off

Blum, Nunes, Prangle, Fearnhead 2012

The error in the ABC approximation can be broken into two parts

1. Choice of summary:

$$\pi(\theta|D) \stackrel{?}{\approx} \pi(\theta|S(D))$$

2. Use of ABC acceptance kernel:

$$\pi(\theta|s_{obs}) \stackrel{?}{\approx} \pi_{ABC}(\theta|s_{obs}) = \int \pi(\theta, s|s_{obs})\mathrm{d}s$$
$$\propto \int \pi_{\epsilon}(s_{obs}|S(x))\pi(x|\theta)\pi(\theta)\mathrm{d}x$$

The first approximation allows the matching between $S(D)$ and $S(X)$ to be done in a lower dimension. There is a trade-off

- $\dim(S)$ small: $\pi(\theta|s_{obs}) \approx \pi_{ABC}(\theta|s_{obs})$, but $\pi(\theta|s_{obs}) \not\approx \pi(\theta|D)$
- $\dim(S)$ large: $\pi(\theta|s_{obs}) \approx \pi(\theta|D)$ but $\pi(\theta|s_{obs}) \not\approx \pi_{ABC}(\theta|s_{obs})$ as curse of dimensionality forces us to use larger ϵ

Choosing summary statistics

If $S(D) = s_{obs}$ is sufficient for θ, i.e., s_{obs} contains all the information contained in D about θ

$$\pi(\theta|s_{obs}) = \pi(\theta|D),$$

then using summaries has no detrimental effect

Choosing summary statistics

If $S(D) = s_{obs}$ is sufficient for θ, i.e., s_{obs} contains all the information contained in D about θ

$$\pi(\theta|s_{obs}) = \pi(\theta|D),$$

then using summaries has no detrimental effect

However, low-dimensional sufficient statistics are rarely available.
How do we choose good low dimensional summaries?

The choice is one of the most important parts of ABC algorithms

Automated summary selection

Blum, Nunes, Prangle and Fearnhead 2012

Suppose we are given a candidate set $\mathcal{S} = (s_1, \ldots, s_p)$ of summaries from which to choose.

Methods break down into groups.

- Best subset selection
  - Joyce and Marjoram 2008
  - Nunes and Balding 2010
- Projection
  - Blum and Francois 2010
  - Fearnhead and Prangle 2012
- Regularisation techniques
  - Blum, Nunes, Prangle and Fearnhead 2012

Best subset selection

Introduce a criterion, e.g,

- τ-sufficiency (Joyce and Marjoram 2008): $s_{1:k-1}$ are τ-sufficient relative to s_k if

$$\begin{aligned}\delta_k &= \sup_\theta \log \pi(s_k|s_{1:k-1},\theta) - \inf_\theta \log \pi(s_k|s_{1:k-1},\theta) \\ &= \text{range}_\theta(\pi(s_{1:k}|\theta) - \pi(s_{1:k-1}|\theta)) \leq \tau\end{aligned}$$

  i.e. adding s_k changes posterior sufficiently.

- Entropy (Nunes and Balding 2010)

Implement within a search algorithm such as forward selection.
Problems:

- assumes every change to posterior is beneficial (see below)
- considerable computational effort required to compute δ_k

Projection: Fearnhead and Prangle 2012

Several statistics from $\mathcal{S}$ may be required to get same info content as a single informative summary.

- project $\mathcal{S}$ onto a lower dimensional highly informative summary vector

The optimal summary statistic (for point estimation) is

$$\tilde{s} = \mathbb{E}(\theta|D)$$

which is usually unknown. They estimate it using the model

$$\theta_i = \mathbb{E}(\theta|D) + e_i = \beta^T f(\mathcal{S}_i) + e_i$$

where $f(\mathcal{S})$ is a vector of functions of $\mathcal{S}$ and $(\theta_i, \mathcal{S}_i)$ are output from a pilot ABC simulation. They choose the set of regressors using, e.g., BIC.

Projection: Fearnhead and Prangle 2012

Several statistics from $\mathcal{S}$ may be required to get same info content as a single informative summary.

- project $\mathcal{S}$ onto a lower dimensional highly informative summary vector

The optimal summary statistic (for point estimation) is

$$\tilde{s} = \mathbb{E}(\theta|D)$$

which is usually unknown. They estimate it using the model

$$\theta_i = \mathbb{E}(\theta|D) + e_i = \beta^T f(\mathcal{S}_i) + e_i$$

where $f(\mathcal{S})$ is a vector of functions of $\mathcal{S}$ and $(\theta_i, \mathcal{S}_i)$ are output from a pilot ABC simulation. They choose the set of regressors using, e.g., BIC.

They then use the single summary statistic

$$\tilde{s} = \hat{\beta}^T f(\mathcal{S})$$

for θ (one summary per parameter).

- Scales well with large p and gives good point estimates.
- Summaries usually lack interpretability and method gives no guarantees about the approximation of the posterior.

Summary warning:

Automated methods are a poor replacement for expert knowledge.

Summary warning:

Automated methods are a poor replacement for expert knowledge. What aspects of the data do we expect our model to be able to reproduce?

- $S(D)$ may be highly informative about θ, but if the model was not built to reproduce $S(D)$ then why should we calibrate to it?
  - many dynamical systems models are designed to model periods and amplitudes, not phase. Summaries that are not phase invariant may still be informative about θ.

  In the case where models and/or priors are mis-specified, this problem can be particularly acute.

Summary warning:

Automated methods are a poor replacement for expert knowledge. What aspects of the data do we expect our model to be able to reproduce?

- $S(D)$ may be highly informative about θ, but if the model was not built to reproduce $S(D)$ then why should we calibrate to it?
  - many dynamical systems models are designed to model periods and amplitudes, not phase. Summaries that are not phase invariant may still be informative about θ.

  In the case where models and/or priors are mis-specified, this problem can be particularly acute.
- The rejection algorithm is usually used in summary selection algorithms, as otherwise we need to rerun the MCMC or SMC sampler for each new summary which is expensive.

Meta-modelling approaches to ABC

Limitations of Monte Carlo methods

Monte Carlo methods are generally guaranteed to succeed if we run them for long enough.

This guarantee is costly and can require more simulation than is possible.

Limitations of Monte Carlo methods

Monte Carlo methods are generally guaranteed to succeed if we run them for long enough.

This guarantee is costly and can require more simulation than is possible.

However,

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

We can use methods that don't suffer in this way, but at the cost of losing the guarantee of success.

Meta-modelling/emulation in deterministic simulators

Sacks *et al.* 1989, Kennedy and O'Hagan 2001

Suppose $f(\theta)$ is a deterministic computer simulator, such as a climate model.

- If $f(\theta)$ is expensive to evaluate, then we can only afford a limited ensemble of simulator evaluations

$$D = \{\theta_i, f(\theta_i)\}_{i=1}^{n}$$

Meta-modelling/emulation in deterministic simulators

Sacks *et al.* 1989, Kennedy and O'Hagan 2001

Suppose $f(\theta)$ is a deterministic computer simulator, such as a climate model.

- If $f(\theta)$ is expensive to evaluate, then we can only afford a limited ensemble of simulator evaluations

$$D = \{\theta_i, f(\theta_i)\}_{i=1}^{n}$$

- We are uncertain about $f(\theta)$ for θ not in the design - code uncertainty.
  - How should we use information in D to do parameter estimation, sensitivity analysis, or prediction?

Meta-modelling/emulation in deterministic simulators

Sacks *et al.* 1989, Kennedy and O'Hagan 2001

Suppose $f(\theta)$ is a deterministic computer simulator, such as a climate model.

- If $f(\theta)$ is expensive to evaluate, then we can only afford a limited ensemble of simulator evaluations

$$D = \{\theta_i, f(\theta_i)\}_{i=1}^{n}$$

- We are uncertain about $f(\theta)$ for θ not in the design - code uncertainty.
  - How should we use information in D to do parameter estimation, sensitivity analysis, or prediction?
- A popular approach is to build an emulator of $f(\cdot)$.

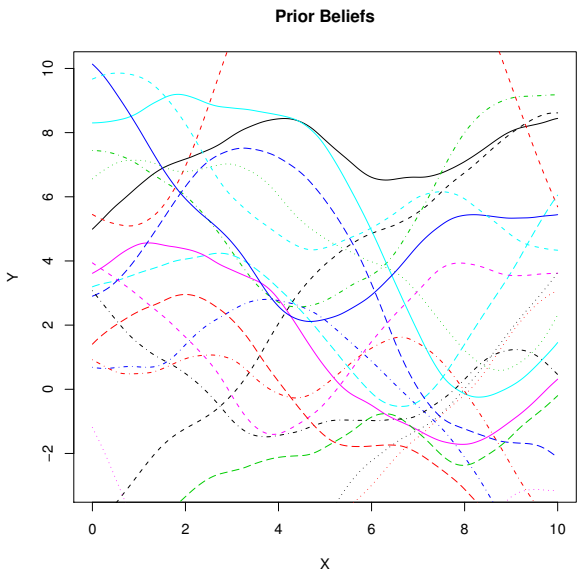
Meta-modelling/emulation for deterministic simulators

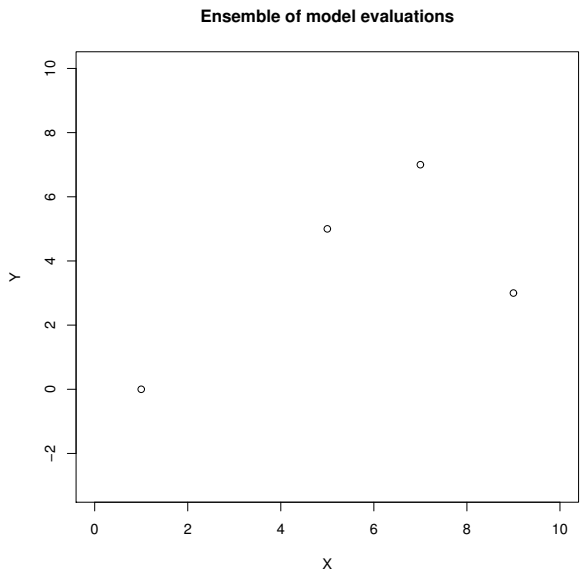
An emulator is a cheap statistical surrogate $\tilde{f}(\theta)$ which approximates $f(\theta)$.

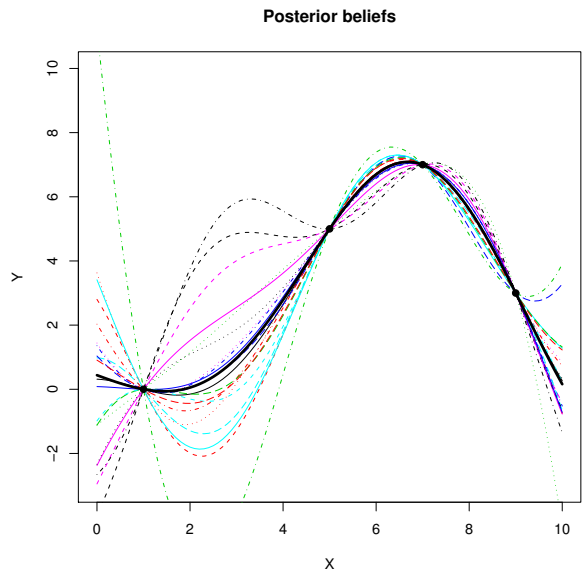
Meta-modelling/emulation for deterministic simulators

An emulator is a cheap statistical surrogate $\tilde{f}(\theta)$ which approximates $f(\theta)$.

Gaussian processes (GP) are a common choice: $\tilde{f}(\cdot) \sim GP(m(\cdot), c(\cdot,\cdot))$



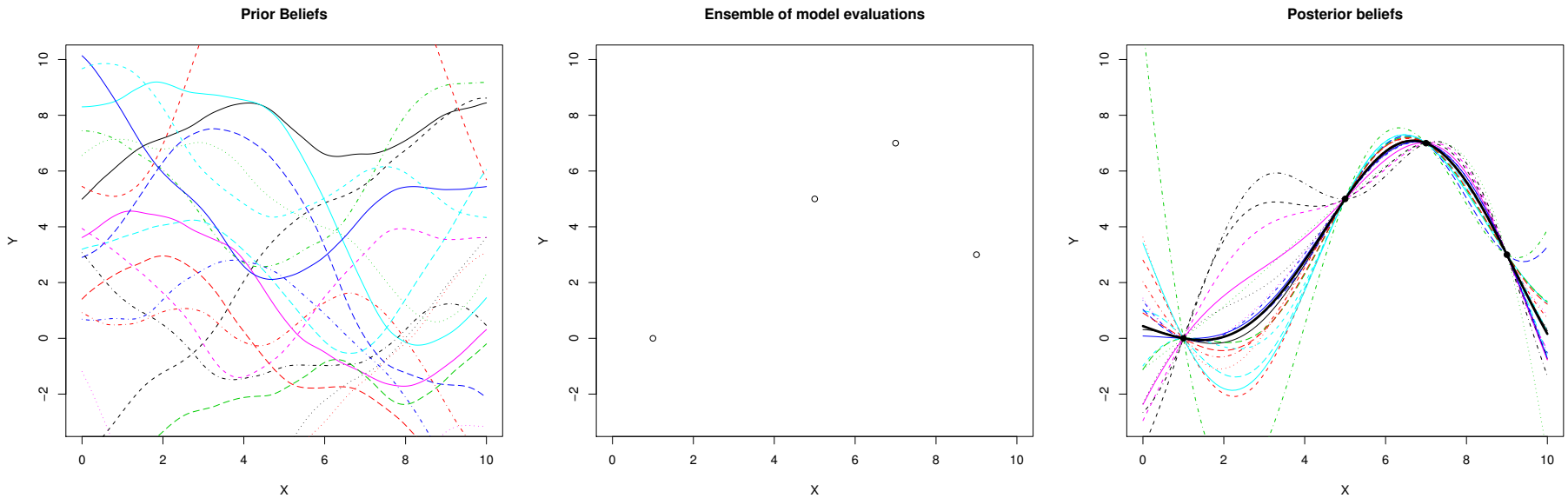




Meta-modelling/emulation for deterministic simulators

An emulator is a cheap statistical surrogate $\tilde{f}(\theta)$ which approximates $f(\theta)$.

Gaussian processes (GP) are a common choice: $\tilde{f}(\cdot) \sim GP(m(\cdot), c(\cdot,\cdot))$



We can then use $\tilde{f}$ in place of f in any analysis.

- GP models include an estimate of their uncertainty
- if trained well, we hope the answer from any statistical analysis incorporates our uncertainty about $f(\cdot)$.

Emulating stochastic models

Cf link to indirect inference (Drovandi, Pettitt, Faddy 2011)

1. Model summaries of the simulator response:
   - e.g., model

     $$m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot,\cdot)) \text{ and } v(\theta) = \mathbb{V}\text{ar}f(\theta) \sim GP(0, c(\cdot,\cdot))$$

     and then assume

     $$f(\theta) \sim N(m(\theta), v(\theta))$$

     Cf. Wood 2010 synthetic likelihood approach.
   - Meeds and Welling 2014, Boukouvalis, Cornford, *et al.* 2009,...

Emulating stochastic models

Cf link to indirect inference (Drovandi, Pettitt, Faddy 2011)

1. Model summaries of the simulator response:
   - e.g., model

     $$m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot,\cdot)) \text{ and } v(\theta) = \mathbb{V}\text{ar}f(\theta) \sim GP(0, c(\cdot,\cdot))$$

     and then assume

     $$f(\theta) \sim N(m(\theta), v(\theta))$$

     Cf. Wood 2010 synthetic likelihood approach.
   - Meeds and Welling 2014, Boukouvalis, Cornford, *et al.* 2009,...
2. Model distribution of simulator output $\pi(f(\theta)|\theta)$, e.g., using Dirichlet process priors (Farah 2011, ...).

Emulating stochastic models

Cf link to indirect inference (Drovandi, Pettitt, Faddy 2011)

1. Model summaries of the simulator response:
   - e.g., model

     $$m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot,\cdot)) \text{ and } v(\theta) = \mathbb{V}\text{ar}f(\theta) \sim GP(0, c(\cdot,\cdot))$$

     and then assume

     $$f(\theta) \sim N(m(\theta), v(\theta))$$

     Cf. Wood 2010 synthetic likelihood approach.
   - Meeds and Welling 2014, Boukouvalis, Cornford, *et al.* 2009,...
2. Model distribution of simulator output $\pi(f(\theta)|\theta)$, e.g., using Dirichlet process priors (Farah 2011, ...).

Disadvantages:

- High dimensional datasets are difficult to model.
- They both involve learning global approximations, i.e. the relationship between D and θ.

Emulating likelihood

Wilkinson 2014, Dahlin and Lindsten 2014

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.

Likelihood estimation

Wilkinson 2013

The GABC framework assumes

$$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{dX}$$
$$\approx \frac{1}{N}\sum \pi(D|X_i)$$

where $X_i \sim \pi(X|\theta)$.

Likelihood estimation

Wilkinson 2013

The GABC framework assumes

$$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{dX}$$
$$\approx \frac{1}{N}\sum \pi(D|X_i)$$

where $X_i \sim \pi(X|\theta)$.

For many problems, we believe the likelihood is continuous and smooth, so that $\pi(D|\theta)$ is similar to $\pi(D|\theta')$ when $\theta - \theta'$ is small

We can model $L(\theta) = \pi(D|\theta)$ and use the model to find the posterior in place of running the simulator.

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.

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

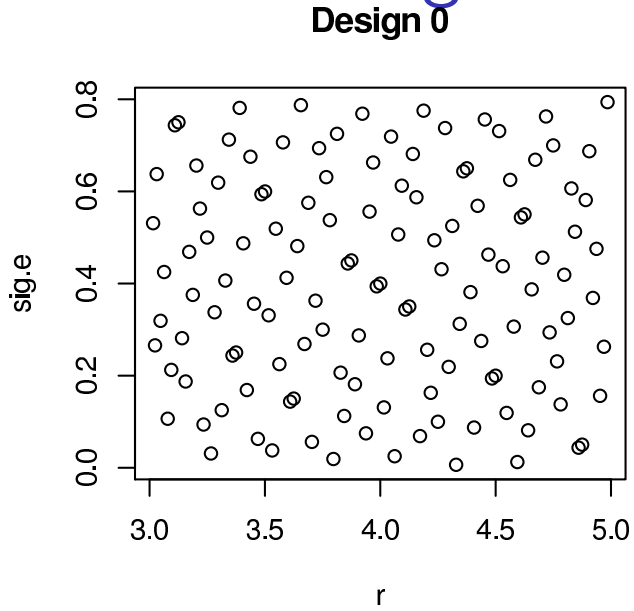
where e_t are independent $N(0, \sigma_e^2)$ process noise

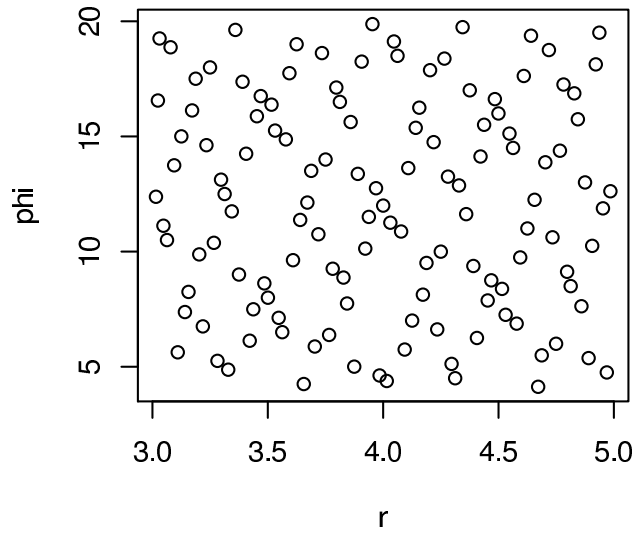
- Assume we observe counts y_t where

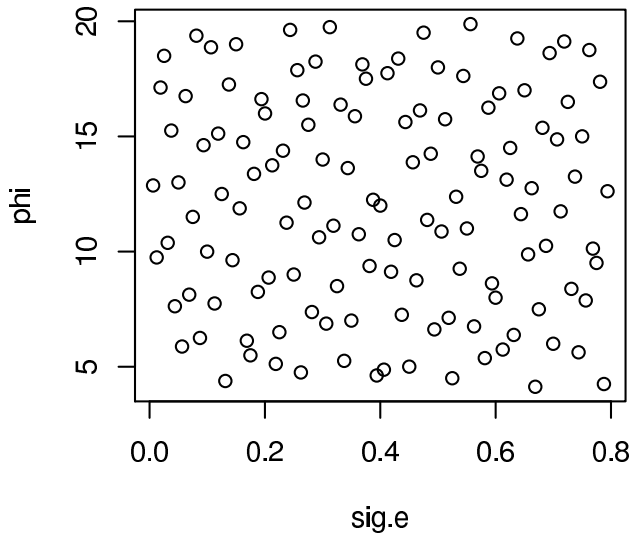
$$y_t \sim Po(\phi N_t)$$

Used in Wood to demonstrate the synthetic likelihood approach.

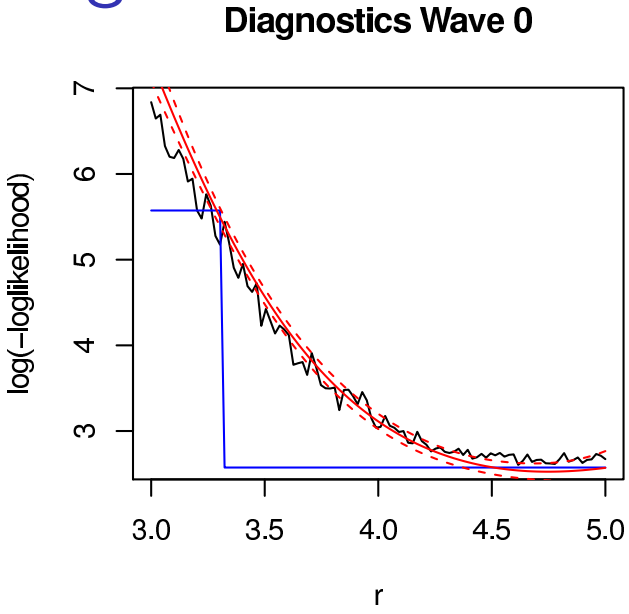
Results - Design 1 - 128 pts

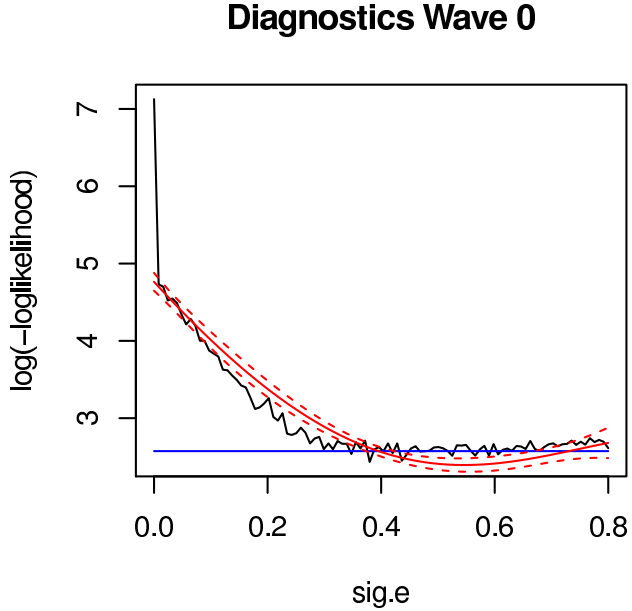


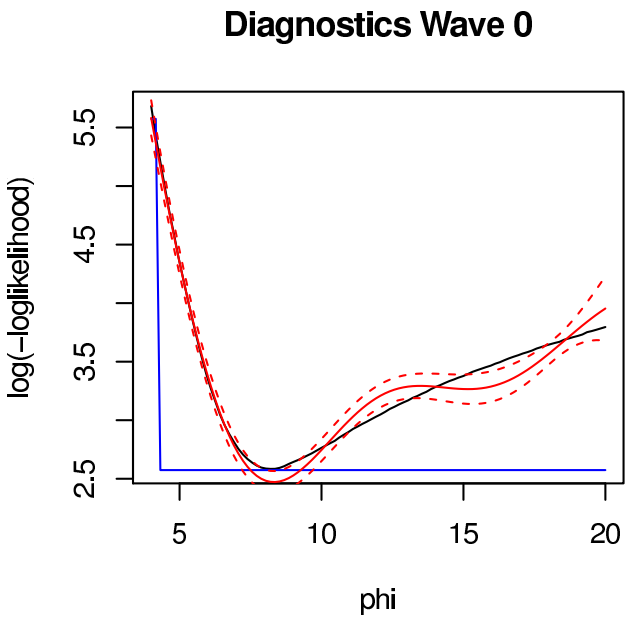




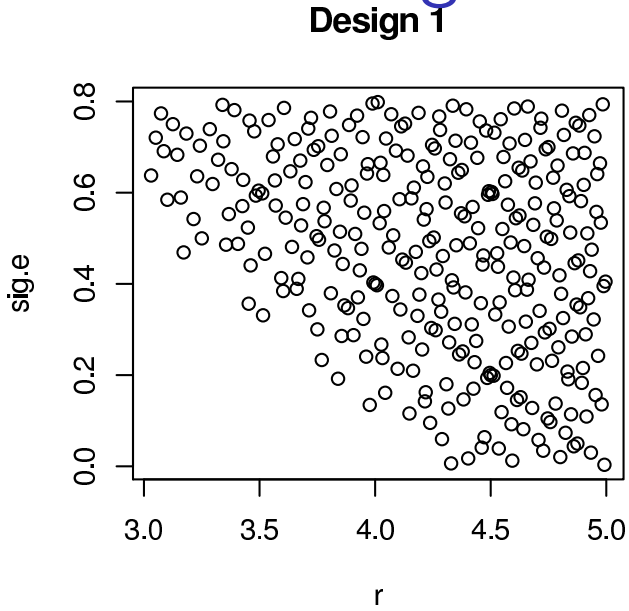
Diagnostics for GP 1 - threshold = 5.6

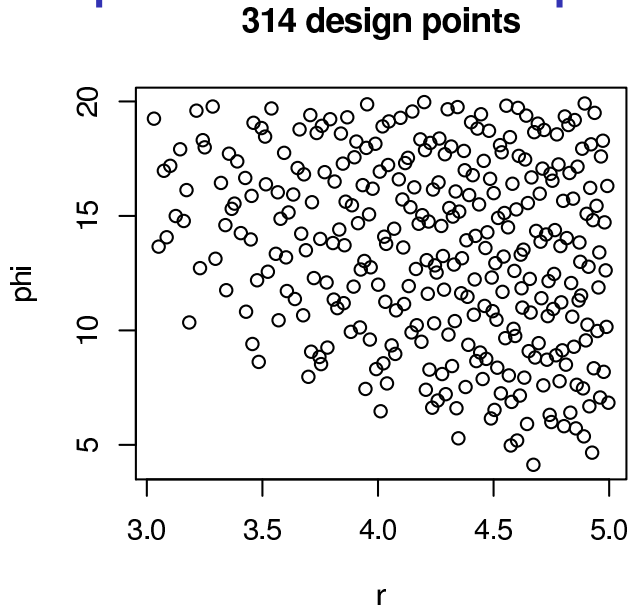


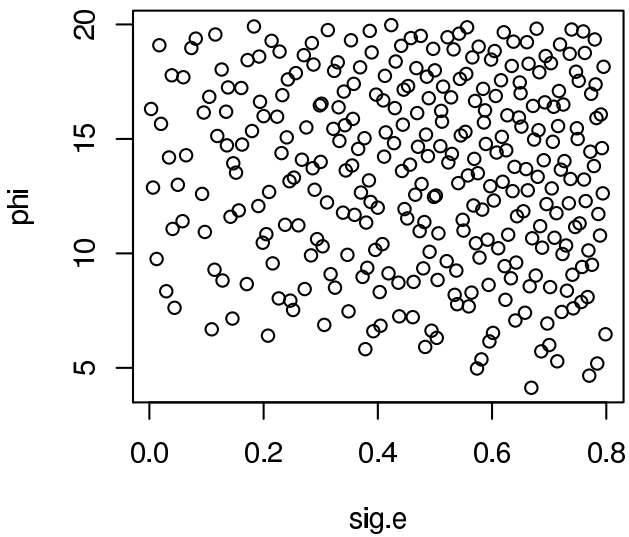




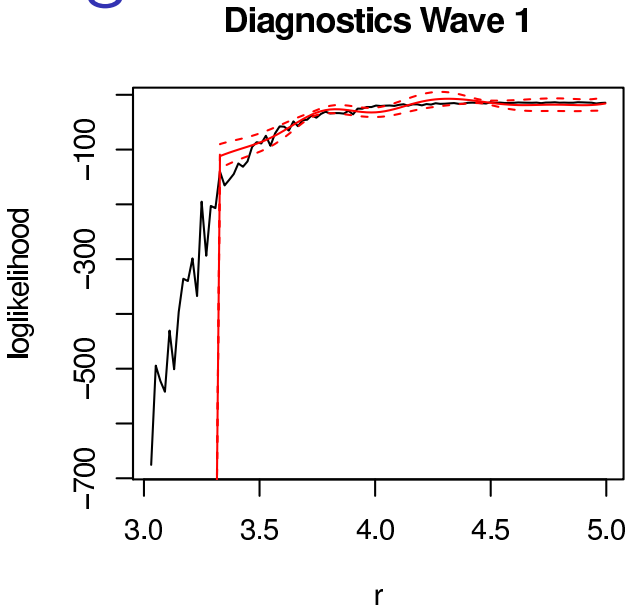
Results - Design 2 - 314 pts - 38% of space implausible

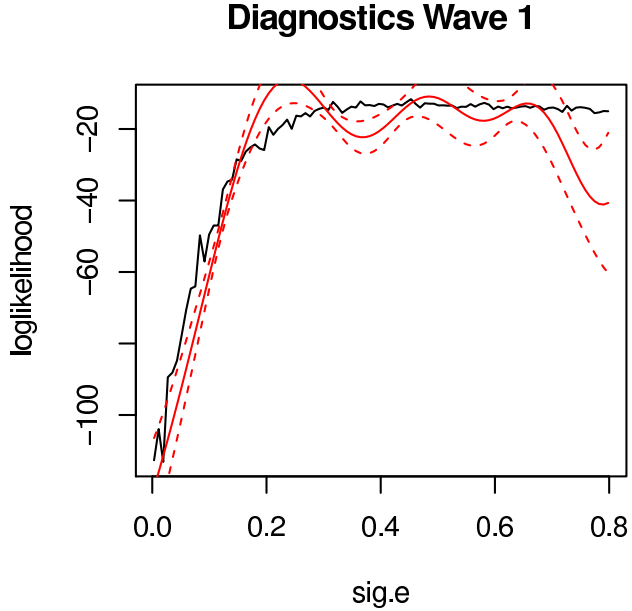


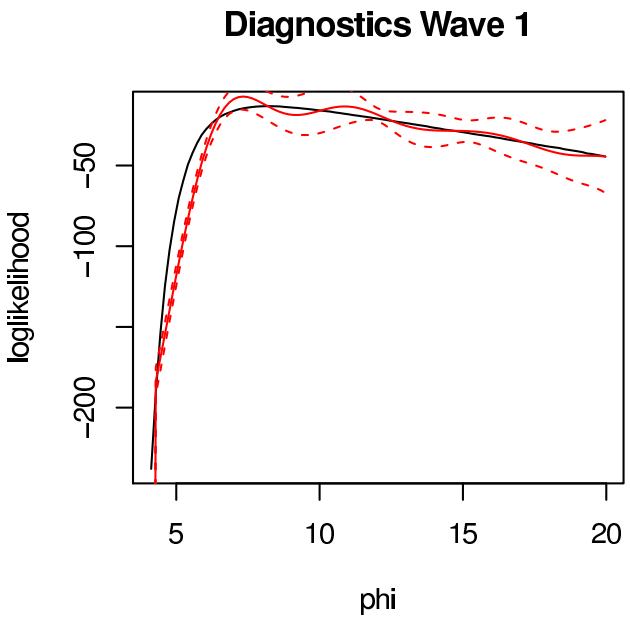




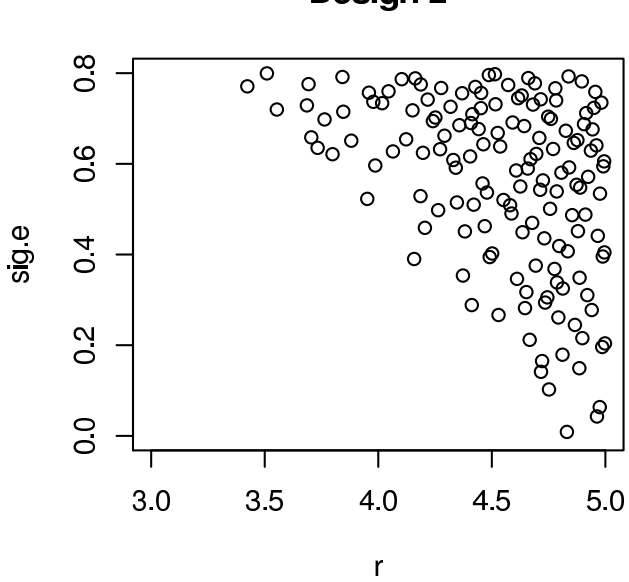
Diagnostics for GP 2 - threshold = -21.8

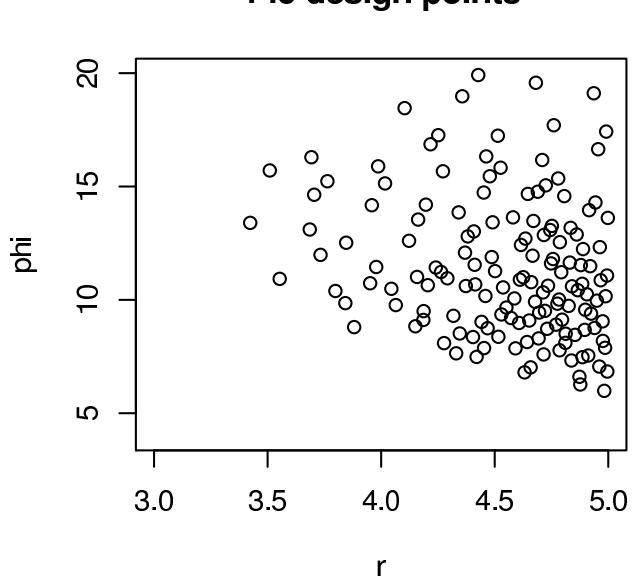


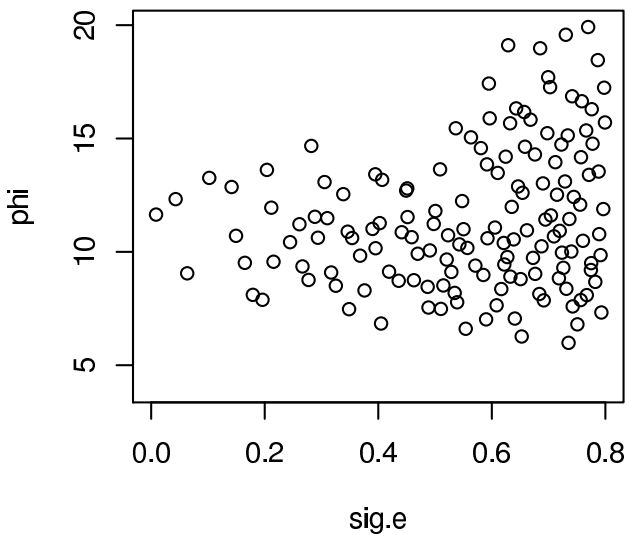




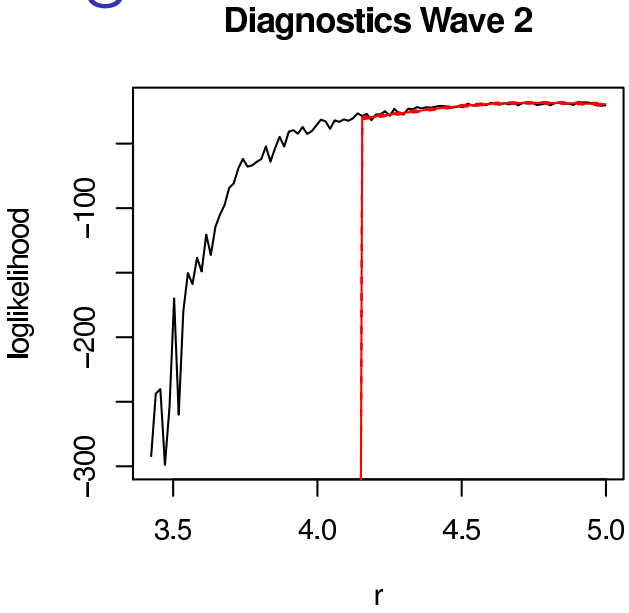
Design 3 - 149 pts - 62% of space implausible

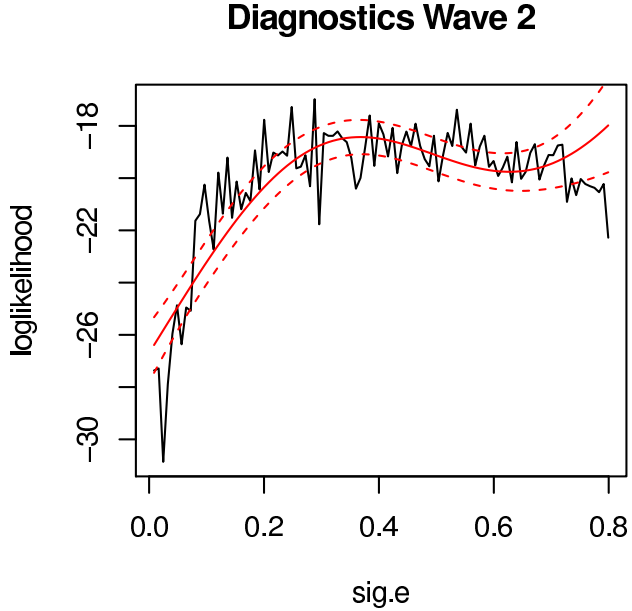


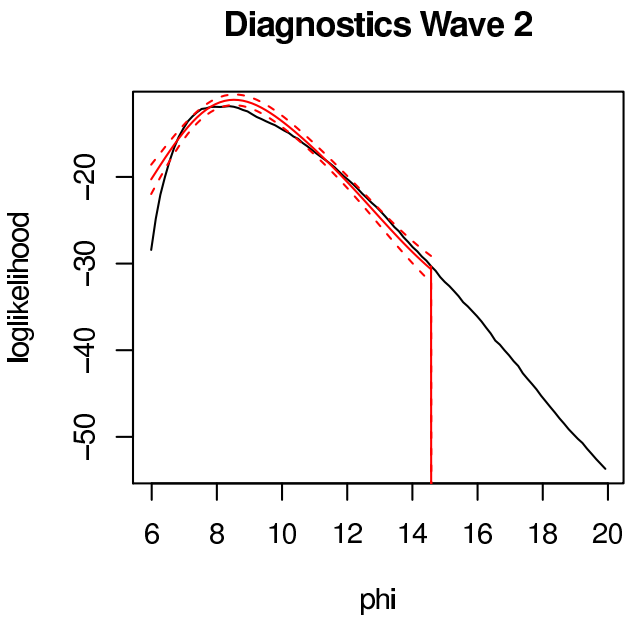




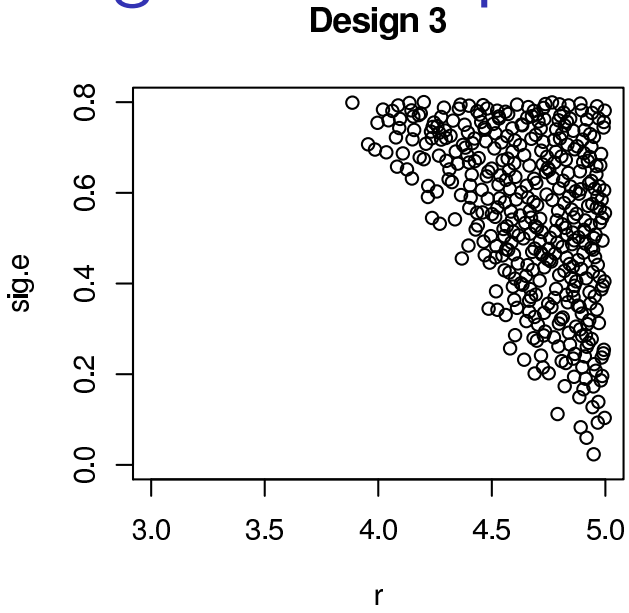
Diagnostics for GP 3 - threshold = -20.7

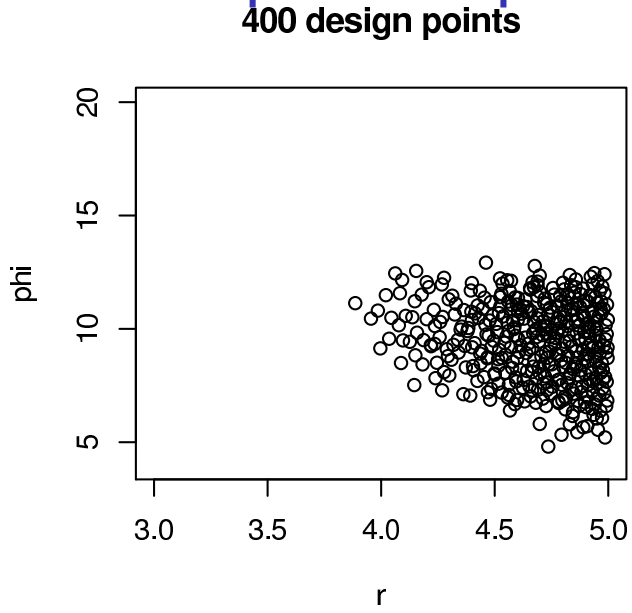


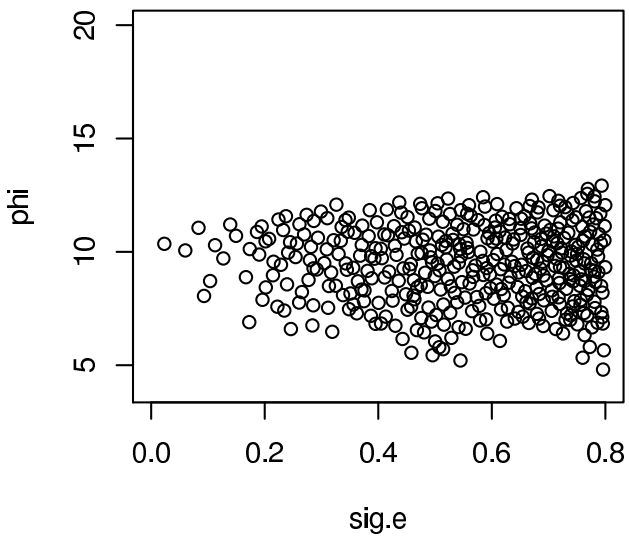




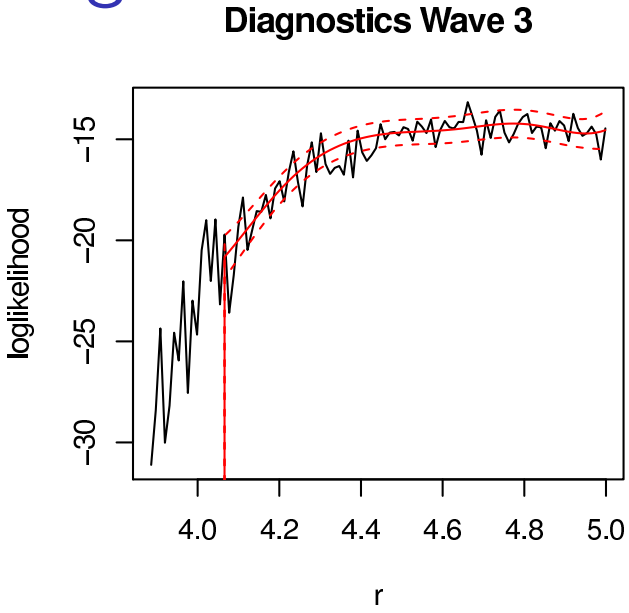
Design 4 - 400 pts - 95% of space implausible

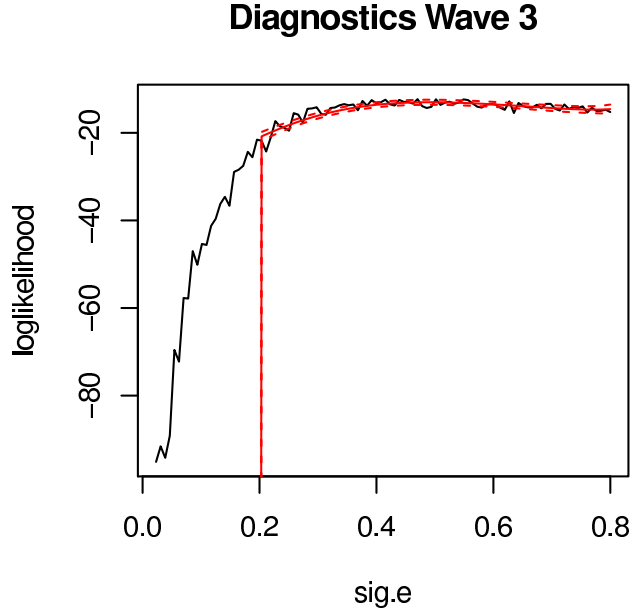


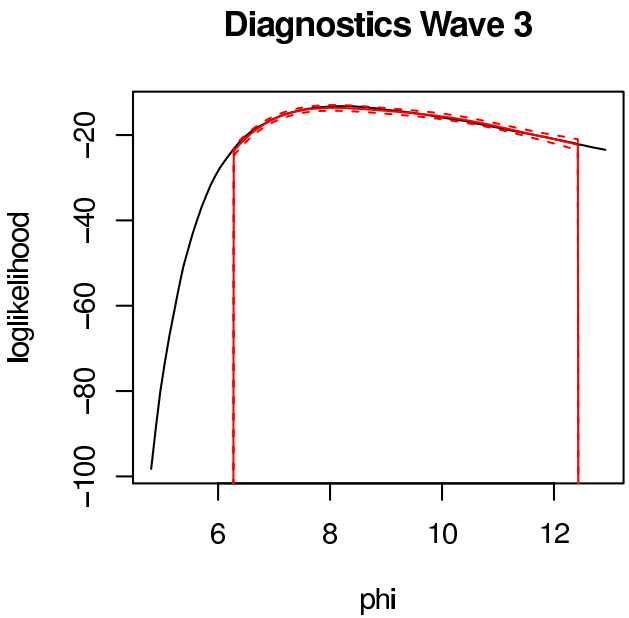




Diagnostics for GP 4 - threshold = -16.4

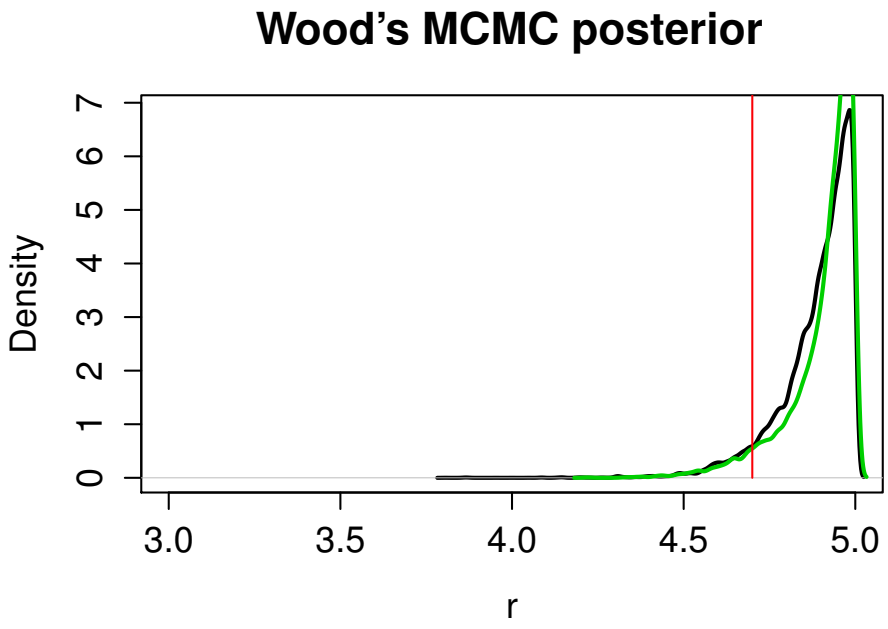


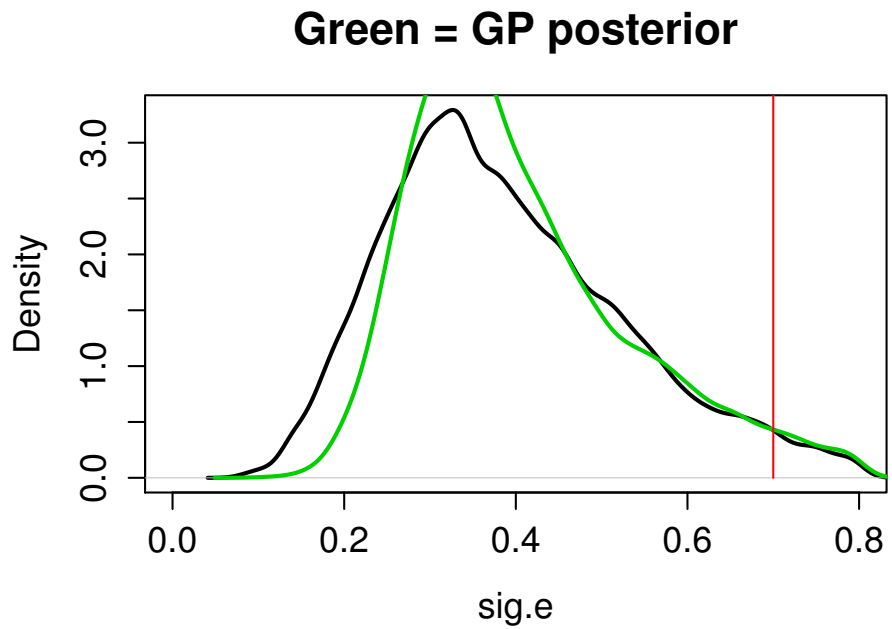


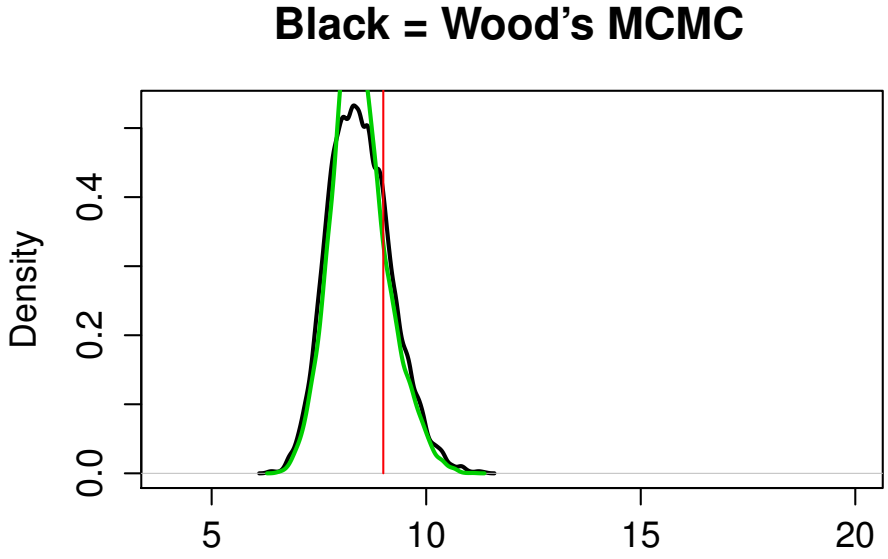


MCMC Results

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.

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.

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.

What's missing:

- Recent machine learning algorithms, e.g., kNN, random forest ...
- Model selection
- ....

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.

What's missing:

- Recent machine learning algorithms, e.g., kNN, random forest ...
- Model selection
- ....

Thank you for listening!

r.d.wilkinson@nottingham.ac.uk
www.maths.nottingham.ac.uk/personal/pmzrdw/

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.