Approximate Bayesian Computation (ABC): inference for intractable computer models

Richard Wilkinson
Data science @LHC

School of Mathematics and Statistics
University of Sheffield

November 10, 2015

An overview of ABC

A taste of several research directions in the ABC community

- Introduction
- Rejection sampling and interpretation
- Efficient algorithms
- Post-hoc regression adjustments
- Surrogate modelling approximations
- Summary statistics
- Model selection

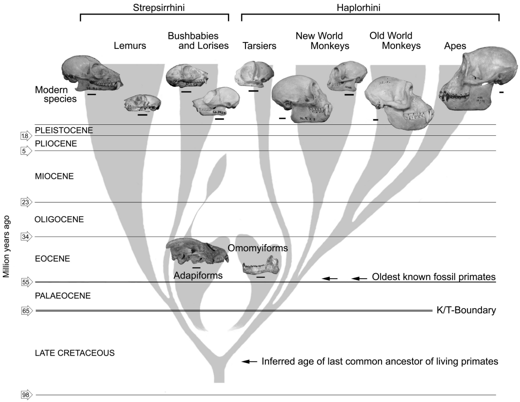
Calibration

- 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 $\pi(D)$.

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)

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

Note that if the lilkelihood is unknown, then we can't find sufficient summary statistics of the data either.

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)

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').

ABC methods are popular in biological disciplines. 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.

'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

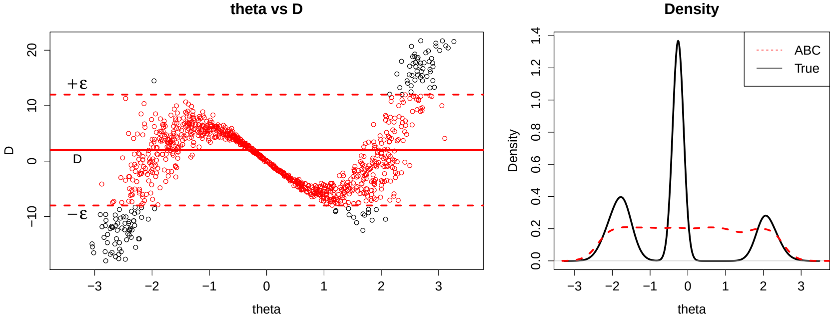
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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

For reasons that will become clear later, we call this *uniform-ABC*.

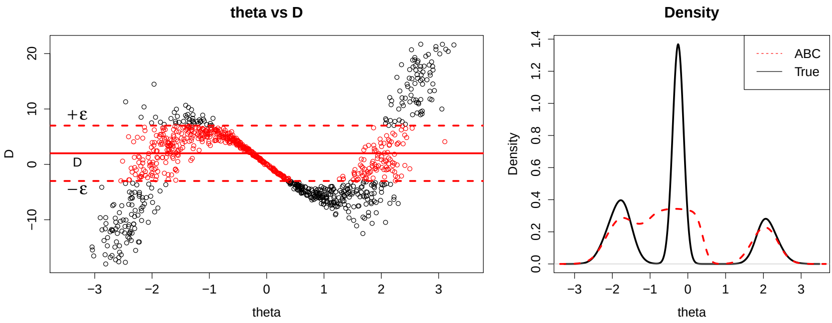
$\epsilon = 10$



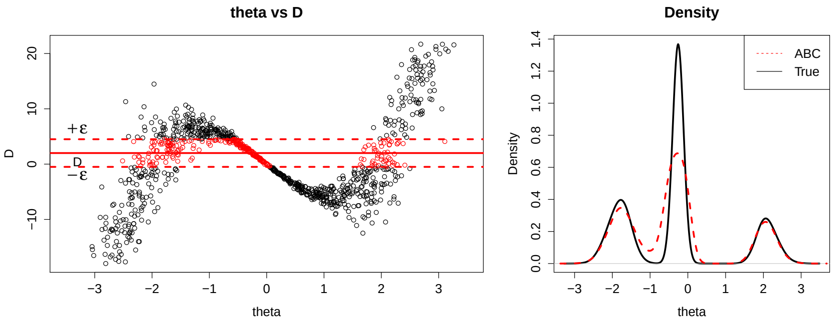
$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta+2)\theta(\theta-2), 0.1+\theta^2)$$

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

$\epsilon = 5$

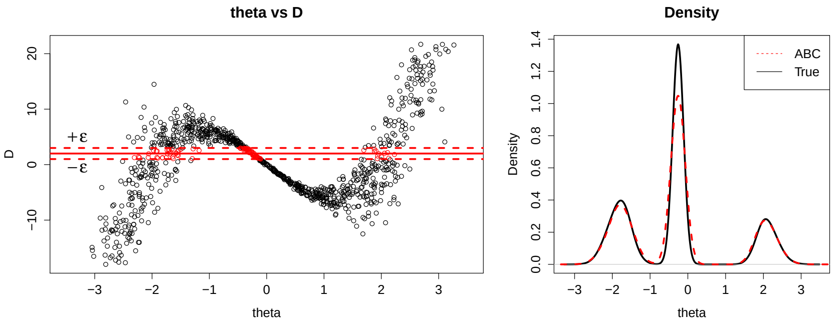

theta vs D
+ε
D
−ε
20
10
0
−10
D
−3
−2
−1
0
1
2
3
theta
Density
ABC
True
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Density
−3
−2
−1
0
1
2
3
theta

$\epsilon = 2.5$



theta vs D
D
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3
theta
Density
Density
ABC
True
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−3
−2
−1
0
1
2
3
theta

$\epsilon = 1$



theta vs D
+ε
−ε
D
theta
−3
−2
−1
0
1
2
3
20
10
0
−10
Density
ABC
True
Density
theta
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Rejection ABC

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

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

Approximate Rejection Algorithm With Summaries

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

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

Rejection ABC

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

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

Approximate Rejection Algorithm With Summaries

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

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

Simple $\rightarrow$ Popular with non-statisticians

ABC as a probability model

W. 2008/2013

We wanted to solve the inverse problem

$$D = f(\theta)$$

but instead ABC solves

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

where the distribution of e depends upon ρ and ϵ.

ABC as a probability model

W. 2008/2013

We wanted to solve the inverse problem

$$D = f(\theta)$$

but instead ABC solves

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

where the distribution of e depends upon ρ and ϵ.

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)

W. 2008, Fearnhead and Prangle 2012

Generalized rejection ABC

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$
2. Accept (θ, X) w.p. $\frac{\pi_\epsilon(D|X)}{\max_x \pi_\epsilon(D|x)}$

where $\pi_\epsilon(D|x)$ is a user specified acceptance kernel, i.e., $P(\text{accept } \theta | f(\theta) = x)$.

Generalized ABC (GABC)

W. 2008, Fearnhead and Prangle 2012

Generalized rejection ABC

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$
2. Accept (θ, X) w.p. $\frac{\pi_\epsilon(D|X)}{\max_x \pi_\epsilon(D|x)}$

where $\pi_\epsilon(D|x)$ is a user specified acceptance kernel, i.e., $P(\text{accept } \theta | f(\theta) = x)$.
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 if $\rho(D, X) \leq \epsilon$

Generalized ABC (GABC)

W. 2008, Fearnhead and Prangle 2012

Generalized rejection ABC

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$
2. Accept (θ, X) w.p. $\frac{\pi_\epsilon(D|X)}{\max_x \pi_\epsilon(D|x)}$

where $\pi_\epsilon(D|x)$ is a user specified acceptance kernel, i.e., $P(\text{accept } \theta | f(\theta) = x)$.
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 if $\rho(D, X) \leq \epsilon$

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

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

Efficient sampling
ABCifying Monte Carlo methods

Rejection is inefficient as it repeatedly samples from prior

More efficient sampling algorithms allow us to spend more time in regions of parameter space likely to lead to accepted values.

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

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

MCMC-ABC

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

Target

$$\pi_{ABC}(\theta, x|D) \propto \mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x|\theta)\pi(\theta)$$

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 as we need the likelihood to cancel in the Metropolis-Hastings (MH) acceptance probability

$$\begin{aligned} r &= \frac{\mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x'|\theta')\pi(\theta')q(\theta',\theta)\pi(x|\theta)}{\mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x|\theta)\pi(\theta)q(\theta,\theta')\pi(x'|\theta')} \\ &= \frac{\mathbb{I}_{\rho(D,x)\leq\epsilon}q(\theta',\theta)\pi(\theta')}{\mathbb{I}_{\rho(D,x)\leq\epsilon}q(\theta,\theta')\pi(\theta)} \end{aligned}$$

MCMC-ABC

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

Target

$$\pi_{ABC}(\theta, x|D) \propto \mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x|\theta)\pi(\theta)$$

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 as we need the likelihood to cancel in the Metropolis-Hastings (MH) acceptance probability

$$\begin{aligned} r &= \frac{\mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x'|\theta')\pi(\theta')q(\theta',\theta)\pi(x|\theta)}{\mathbb{I}_{\rho(D,x)\leq\epsilon}\pi(x|\theta)\pi(\theta)q(\theta,\theta')\pi(x'|\theta')} \\ &= \frac{\mathbb{I}_{\rho(D,x)\leq\epsilon}q(\theta',\theta)\pi(\theta')}{\mathbb{I}_{\rho(D,x)\leq\epsilon}q(\theta,\theta')\pi(\theta)} \end{aligned}$$

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.

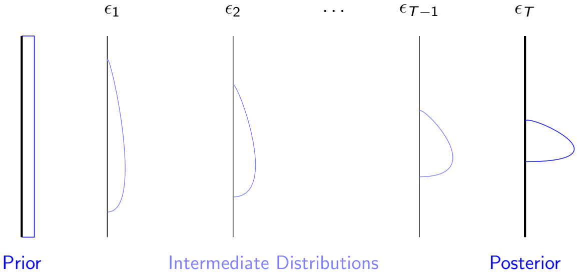
Sequential ABC algorithms

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 sequential methods.

Choose upon a sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ and let π_t be the ABC approximation when we using tolerance ϵ_t.
We aim to sample N particles successively from

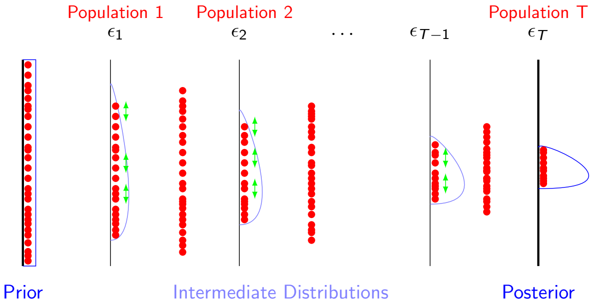
$$\pi_1(\theta), \ \ldots, \ \pi_T(\theta) = \text{target}$$



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

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

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



Picture from Toni and Stumpf 2010 tutorial

Regression Adjustment

References:

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

Post-hoc regression adjustments

Beaumont *et al.* 2002, Blum and Francois 2010

Consider the relationship between the conditional expectation of θ and s:

$$\mathbb{E}(\theta|s) := m(s)$$

Think of this as a model for the conditional density $\pi(\theta|s)$: for fixed s

$$\theta_i = m(s) + e_i$$

where $\theta_i \sim \pi(\theta|s)$ and e_i are zero-mean and uncorrelated

Post-hoc regression adjustments

Beaumont *et al.* 2002, Blum and Francois 2010

Consider the relationship between the conditional expectation of θ and s:

$$\mathbb{E}(\theta|s) := m(s)$$

Think of this as a model for the conditional density $\pi(\theta|s)$: for fixed s

$$\theta_i = m(s) + e_i$$

where $\theta_i \sim \pi(\theta|s)$ and e_i are zero-mean and uncorrelated

Suppose we've estimated $m(s)$ by $\widehat{m}(s)$ from samples $\{\theta_i, s_i\}$.
Estimate the posterior mean by

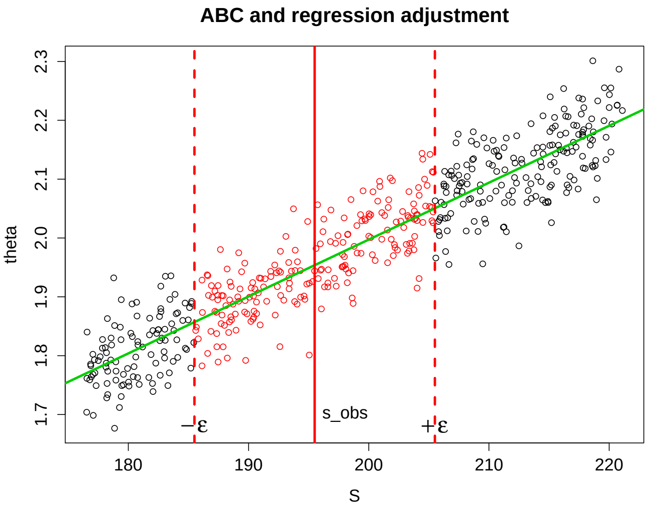
$$\mathbb{E}(\theta|s_{obs}) \approx \widehat{m}(s_{obs}),$$

and form the empirical residuals

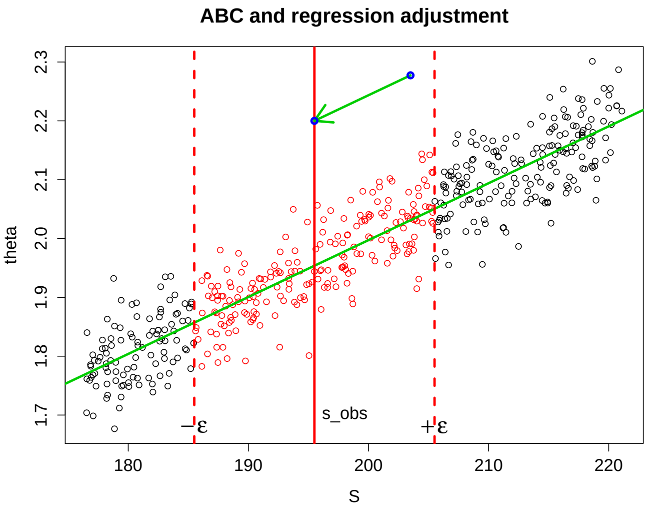
$$\hat{e}_i = \theta_i - \widehat{m}(s_i)$$

We can approximate the posterior $\pi(\theta|s_{obs})$ by adjusting the parameters

$$\theta_i^* = \widehat{m}(s_{obs}) + \hat{e}_i = \theta_i + (\widehat{m}(s_{obs}) - \widehat{m}(s_i))$$

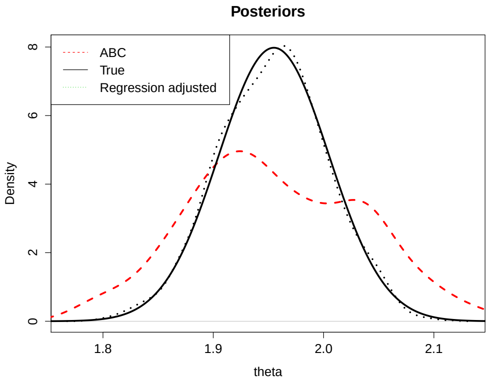


In rejection ABC, the red points are used to approximate the histogram.



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.

Normal-normal conjugate model, linear regression



Regression-adjusted posterior more confident, as the θ_i have been adjusted to account for the discrepancy between s_i and s_{obs}

- Allows larger ϵ for same accuracy
- Sequential algorithms can not easily be adapted, thus regression adjustment used with rejection sampling only.

Surrogate/emulator methods

References:

- Kennedy and O'Hagan 2001
- Wilkinson 2014
- Conrad, Marzouk, Pillai, Smith 2014
- Meeds and Welling 2015
- Corrander *et al.* 2015

Surrogate/emulator approximations

Sacks *et al.* 1989, Kennedy *et al.* 2001, W. 2014/15, Meeds *et al.* 2015, Corrander *et al.* 2015

ABC requires a large number of simulator runs:

- Suppose 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

Surrogate/emulator approximations

Sacks *et al.* 1989, Kennedy *et al.* 2001, W. 2014/15, Meeds *et al.* 2015, Corrander *et al.* 2015

ABC requires a large number of simulator runs:

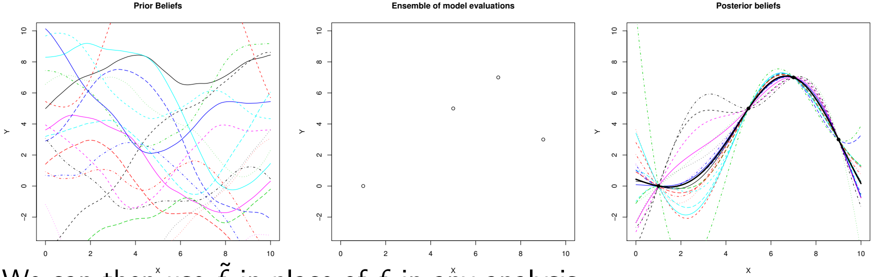
- Suppose we can only afford a limited ensemble of simulator evaluations

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

- We are uncertain about $f(\theta)$ for θ not in the design

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.

Likelihood estimation

W. 2013

It can be shown that ABC replaces the true likelihood $L(\theta) \equiv \pi(D|\theta)$ by an ABC likelihood

$$L_{\text{ABC}}(\theta) = \int \mathbb{I}_{\rho(D,X) \leq \epsilon} \pi(X|\theta) \mathrm{dX}$$

We can estimate this using repeated runs from the simulator

$$\hat{L}_{\text{ABC}}(\theta) \approx \frac{1}{N} \sum \mathbb{I}_{\rho(D,X_i) \leq \epsilon}$$

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

Likelihood estimation

W. 2013

It can be shown that ABC replaces the true likelihood $L(\theta) \equiv \pi(D|\theta)$ by an ABC likelihood

$$L_{\text{ABC}}(\theta) = \int \mathbb{I}_{\rho(D,X)\leq\epsilon}\pi(X|\theta)\mathrm{dX}$$

We can estimate this using repeated runs from the simulator

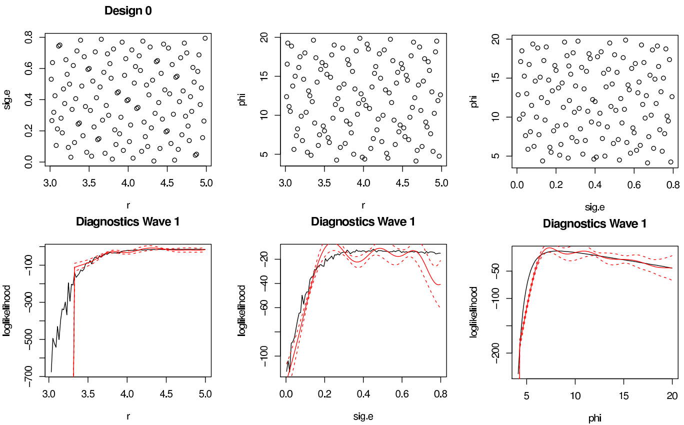
$$\hat{L}_{\text{ABC}}(\theta) \approx \frac{1}{N}\sum \mathbb{I}_{\rho(D,X_i)\leq\epsilon}$$

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

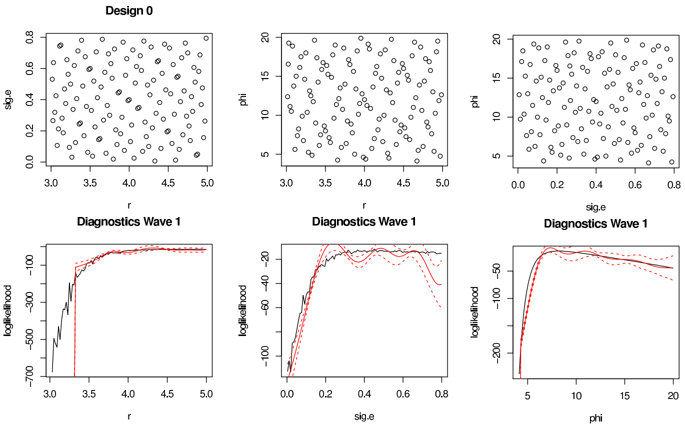
For many problems, we believe the likelihood is continuous and smooth, so that $L_{\text{ABC}}(\theta)$ is similar to $L_{\text{ABC}}(\theta')$ when $\theta - \theta'$ is small

We can model $L_{\text{ABC}}(\theta)$ and use the model to find the posterior in place of running the simulator.

Ricker Model



Ricker Model



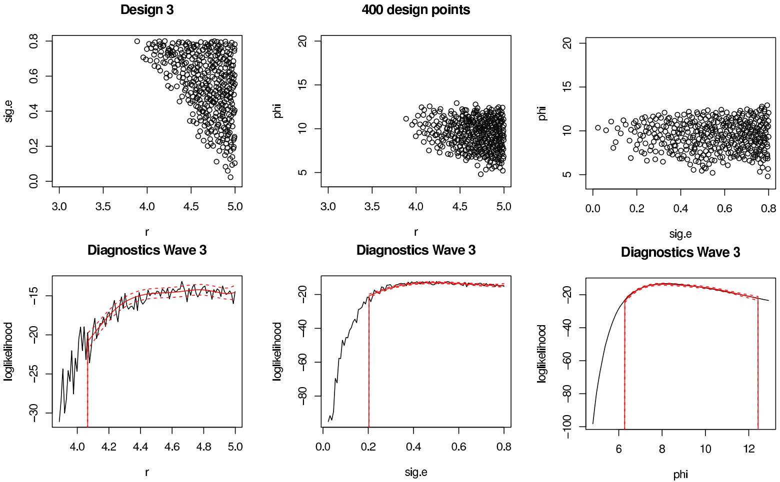
It is usually too difficult to model $L(\theta)$ for all θ

- Sufficient to know $L(\theta)$ in regions of high likelihood, and to know that it is small elsewhere.

Use this initial model to rule out large parts of parameter space as implausible using a conservative heuristic.

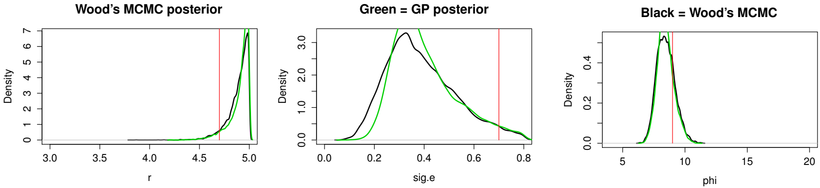
Build a better model, repeat.

Ricker Model - third wave



MCMC Results

Comparison with Wood 2010, synthetic likelihood approach



- The Wood 2010 ABC-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, over 98% of the original input space was ruled out as implausible

- the MCMC sampler does not waste time exploring those regions.

Implausibility

When using emulators for history-matching and ABC, we want to estimate

$$p(\theta) = \mathbb{P}(\text{Accept } \theta)$$

based upon a GP model of the simulator or likelihood

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

Implausibility

When using emulators for history-matching and ABC, we want to estimate

$$p(\theta) = \mathbb{P}(\text{Accept } \theta)$$

based upon a GP model of the simulator or likelihood

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

The key determinant of emulator accuracy is the design used

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

Usual design choices are space filling designs

- Maximin latin hypercubes, Sobol sequences

Entropic designs

Active learning/sequential design

However, space filling designs are good for global approximations, but wasteful for calibration

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

$$p(\theta) = \mathbb{P}(\text{Accept } \theta | D_n)$$

to guide the choice of design points

Entropic designs

Active learning/sequential design

However, space filling designs are good for global approximations, but wasteful for calibration

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

$$p(\theta) = \mathbb{P}(\text{Accept } \theta | D_n)$$

to guide the choice of design points

The entropy of the classification surface is

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

One (unwise) approach is to choose the next design point where we are most uncertain.

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

- design points tend to accumulate on the edge of the domain Θ.

Expected average entropy

Chevalier *et al.* 2014

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

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

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

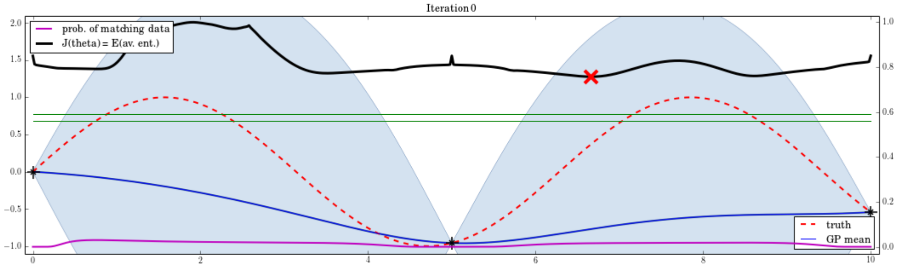
- Choose the next design point, θ_{n+1}, to minimise the expected average entropy

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

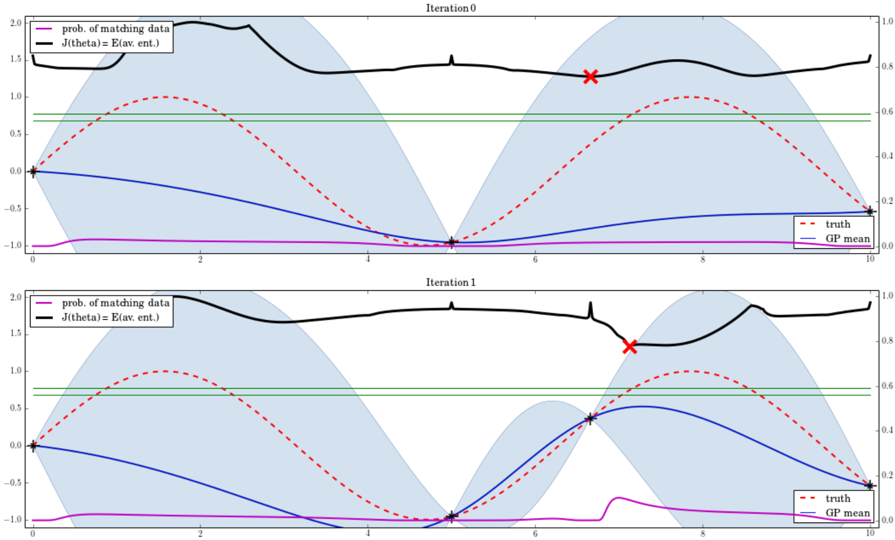
  where

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

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

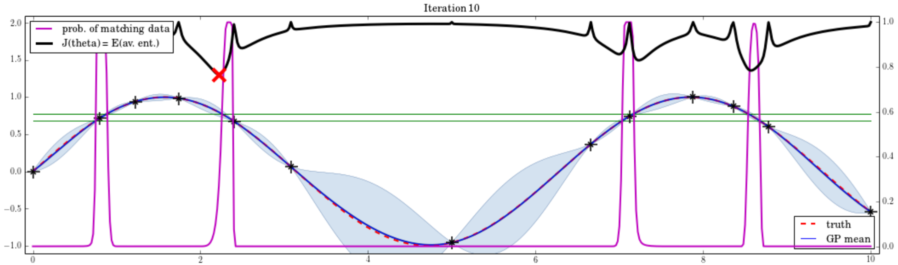


Toy 1d example $f(\theta) = \sin\theta$ - Expected entropy



Toy 1d: min expected entropy vs max entropy

After 10 iterations, choosing the point of maximum entropy



we have found the plausible region to reasonable accuracy.

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

Summary statistic selection: error trade-off

Fearnhead and Prangle 2012, Blum, Nunes, Prangle, Fearnhead 2012

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

1. Choice of summary:

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

Summary statistic selection: error trade-off

Fearnhead and Prangle 2012, Blum, Nunes, Prangle, Fearnhead 2012

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

1. Choice of summary:
$$\pi(\theta|D) \overset{?}{\approx} \pi(\theta|S(D))$$
2. Use of ABC acceptance kernel:
$$\begin{aligned}\pi(\theta|s_{obs}) \overset{?}{\approx} \pi_{ABC}(\theta|s_{obs}) &= \int \pi(\theta, s|s_{obs})\mathrm{d}s \\ &\propto \int \mathbb{I}_{\rho(s_{obs}, S(X))\leq\epsilon}\pi(x|\theta)\pi(\theta)\mathrm{d}x\end{aligned}$$

Summary statistic selection: error trade-off

Fearnhead and Prangle 2012, 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:

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

$$\propto \int \mathbb{I}_{\rho(s_{obs}, S(X)) \leq \epsilon} \pi(x|\theta)\pi(\theta)\mathrm{d}x$$

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

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
  - Pudlo, Marin, Estoup, Cornuet, Gautier, Robert 2014.
- Regularisation techniques
  - Blum, Nunes, Prangle and Fearnhead 2012

Machine learning type tools increasingly used to find good discriminating summary statistics.

Model selection

Model selection

W. 2007, Grelaud *et al.* 2009

But often we want to compare models $\rightarrow$ Bayes factors

$$B_{12} = \frac{\pi(D|M_1)}{\pi(D|M_2)}$$

where $\pi(D|M_i) = \int \mathbb{I}_{\rho(D,X)\leq\epsilon}\pi(x|\theta, M_i)\pi(\theta)\mathrm{d}x\mathrm{d}\theta$.

Model selection

W. 2007, Grelaud *et al.* 2009

But often we want to compare models $\rightarrow$ Bayes factors

$$B_{12} = \frac{\pi(D|M_1)}{\pi(D|M_2)}$$

where $\pi(D|M_i) = \int \mathbb{I}_{\rho(D,X)\leq\epsilon}\pi(x|\theta, M_i)\pi(\theta)\mathrm{d}x\mathrm{d}\theta$.

For rejection ABC

$$\pi(D|M) \approx \frac{1}{N}\sum \mathbb{I}_{\rho(D,X_i)\leq\epsilon}$$

where $X_i \sim M(\theta_i)$ with $\theta_i \sim \pi(\theta)$.

Summary statistics for model selection

Didelot *et al.* 2011, Robert *et al.* 2011

Care needs to be taken with regard summary statistics for model selection. Everything is okay if we target

$$B_S = \frac{\pi(S(D)|M_1)}{\pi(S(D)|M_2)}$$

Then the ABC estimator $\hat{B}_S^\epsilon \to B_S$ as $\epsilon \to 0, N \to \infty$ (Didelot *et al.* 2011).

Summary statistics for model selection

Didelot *et al.* 2011, Robert *et al.* 2011

Care needs to be taken with regard summary statistics for model selection. Everything is okay if we target

$$B_S = \frac{\pi(S(D)|M_1)}{\pi(S(D)|M_2)}$$

Then the ABC estimator $\hat{B}_S^\epsilon \to B_S$ as $\epsilon \to 0, N \to \infty$ (Didelot *et al.* 2011).
However,

$$\frac{\pi(S(D)|M_1)}{\pi(S(D)|M_2)} \neq \frac{\pi(D|M_1)}{\pi(D|M_2)} = B_D$$

even if S is a sufficient statistic!

S sufficient for $f_1(D|\theta_1)$ and $f_2(D|\theta_2)$ does not imply sufficiency for $\{m, f_m(D|\theta_m)\}$. Hence $\hat{B}_S^\epsilon \not\to B_D$.
Not a problem if we view inference as conditional on a carefully chosen S.

Conclusions

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

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

The main challenges for ABC are

- finding good summary statistics for high dimensional problems
- dealing with computationally expensive simulators.

Conclusions

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

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

The main challenges for ABC are

- finding good summary statistics for high dimensional problems
- dealing with computationally expensive simulators.

Thank you for listening!