‘UQ’ perspectives on ABC (approximate Bayesian computation)

Richard Wilkinson

School of Maths and Statistics
University of Sheffield

January 12, 2018

Inverse problems/Calibration/Parameter estimation/...

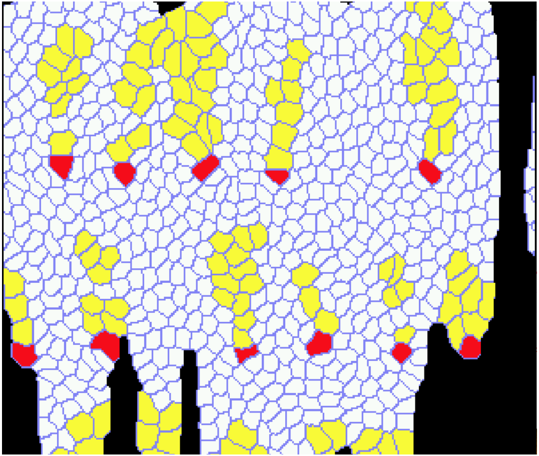
- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- The inverse-problem: observe data D, estimate parameter values θ which explain the data.

The Bayesian approach is to find the posterior distribution

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

posterior $\propto$

prior × likelihood



Introduction

Simulation from Andrea Sottoriva

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.
- want to infer number of stem cells by comparing patterns with real data

Each simulation takes $\sim$ 1 hour

There are plenty of stochastic models which

- have unknown parameters
- are stochastic
- have unknown likelihood function
- are computationally expensive
- are imperfect

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 widely used in several scientific disciplines (particularly comp bio + genetics), and has similarities with history-matching. They are

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

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. Regression adjustments/ post-hoc corrections

iv. Summary statistics

v. Accelerating ABC using meta-models

vi. Inference for misspecified 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:

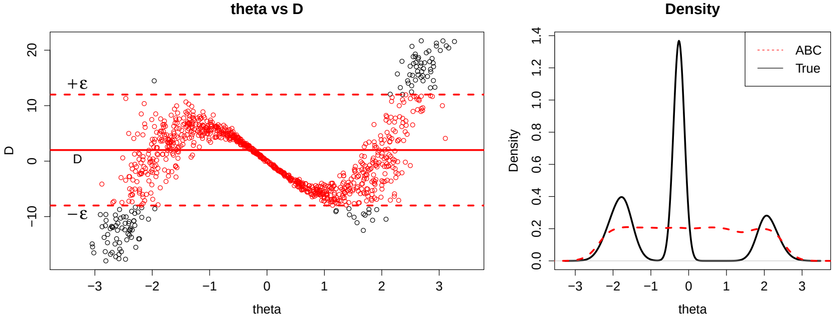
Uniform Rejection Algorithm

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

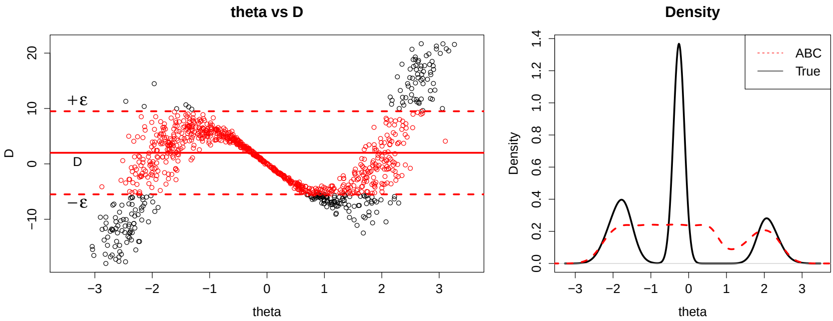
$\epsilon = 10$



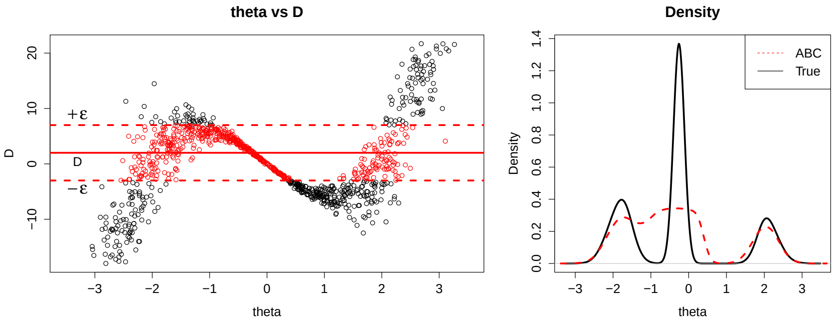
$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 7.5$

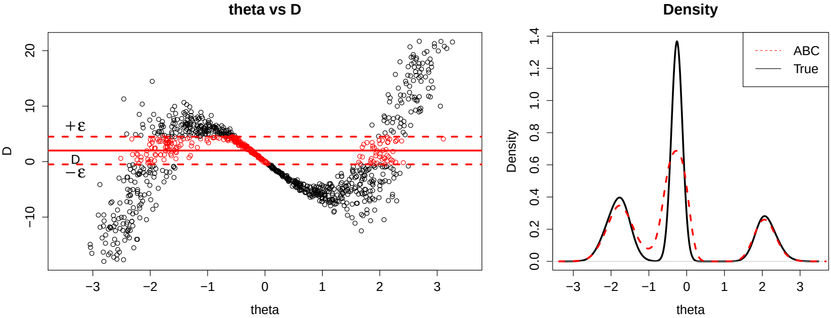

theta vs D
D
20
10
0
−10
+ε
D
−ε
−3
−2
−1
0
1
2
3
theta
Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
ABC
True
−3
−2
−1
0
1
2
3
theta

$\epsilon = 5$



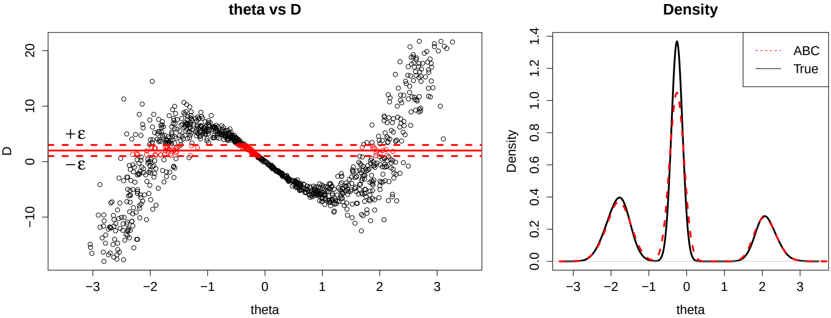
theta vs D
D
20
10
0
−10
+ε
D
−ε
−3
−2
−1
0
1
2
3
theta
Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
ABC
True
−3
−2
−1
0
1
2
3
theta

$\epsilon = 2.5$



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

$\epsilon = 1$



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

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 → Popular with non-statisticians

ABC as a probability model

W. 2008/13

We wanted to solve the inverse problem

$$D = f(\theta)$$

but instead ABC solves

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

ABC as a probability model

W. 2008/13

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)

W. 2008/13

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)

W. 2008/13

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

Scoring

- The tolerance ϵ, distance ρ, summary $S(D)$ (or variations thereof) determine the theoretical 'accuracy' of the approximation
-

Computation

- Computing the approximate posterior for any given score is usually hard.
- There is a trade-off between accuracy achievable in the approximation (size of ϵ), and the information loss incurred when summarizing

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

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

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

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

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

Regression Adjustment

References:

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

Regression Adjustment

Beaumont *et al.* 2002

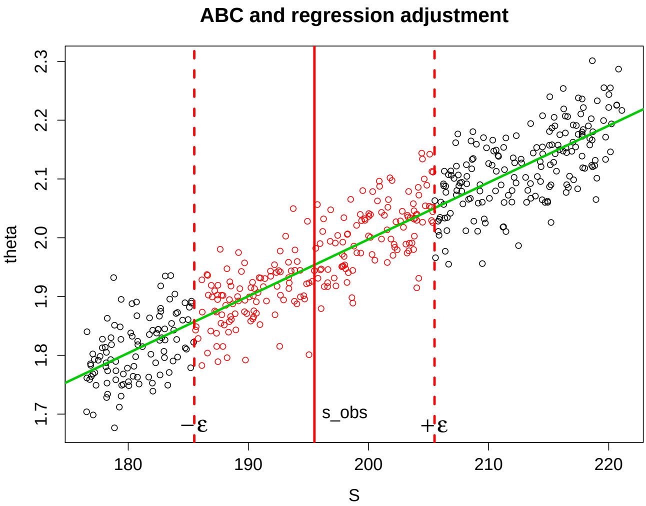
Post-hoc adjustment of the parameter values to try to weaken the effect of the discrepancy between $S(X) = s$ and $S(D) = s_{obs}$ is often used as an alternative to efficient sampling

Two key ideas

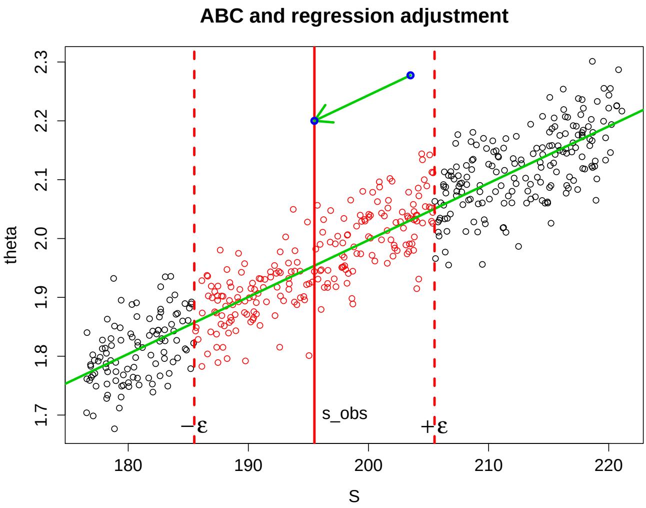
- use non-parametric kernel density estimation to emphasise the best simulations
- learn a non-linear model for the conditional expectation $\mathbb{E}(\theta|s)$ as a function of s and use this to learn the posterior at s_{obs}.

Allows us to use a larger tolerance, and can substantially improve posterior accuracy.

Sequential algorithms (MCMC, SMC etc) can not easily be adapted, and so only used with simple rejection sampling.



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

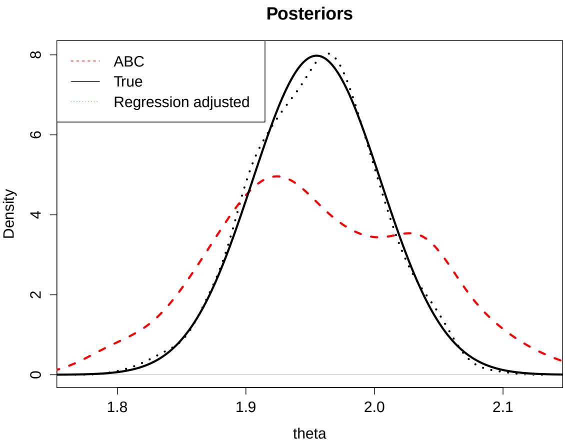
$$\sum (\theta_i - m(s_i))^2 K_\epsilon(s_i - s_{obs})$$

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



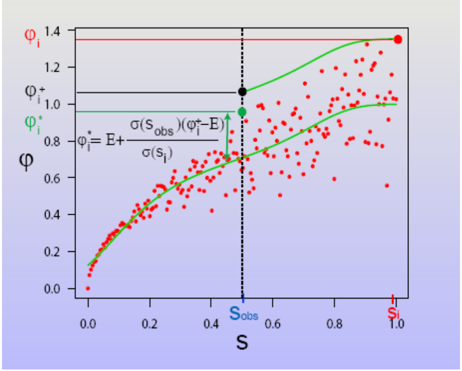
The same 200 data points in both approximations. The regression-adjusted ABC gives a more confident posterior, as the θ_i have been adjusted to account for the discrepancy between s_i and s_{obs}

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 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

Choosing summary statistics

Blum, Nunes, Prangle, Fearnhead 2012

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

Blum, Nunes, Prangle, Fearnhead 2012

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?

Choosing summary statistics

Blum, Nunes, Prangle, Fearnhead 2012

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?
Warning: automated methods are a poor replacement for expert knowledge.
Instead ask what aspects of the data do we expect our model to be able to reproduce?

- $S(D)$ may be highly restrictive about θ, but not necessarily informative, particular if the model is mis-specified.

Error trade-off

Fearnhead and Prangle 2012

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

1. Choice of summary:

$$\pi(\theta|D) \overset{?}{\approx} \pi(\theta|s_{obs})$$

Error trade-off

Fearnhead and Prangle 2012

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

1. Choice of summary:
$$\pi(\theta|D) \overset{?}{\approx} \pi(\theta|s_{obs})$$

2. Use of ABC acceptance kernel:
$$\pi(\theta|s_{obs}) \overset{?}{\approx} \pi_{ABC}(\theta|s_{obs})$$

Error trade-off

Fearnhead and Prangle 2012

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

1. Choice of summary:
$$\pi(\theta|D) \overset{?}{\approx} \pi(\theta|s_{obs})$$
2. Use of ABC acceptance kernel:
$$\pi(\theta|s_{obs}) \overset{?}{\approx} \pi_{ABC}(\theta|s_{obs})$$

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

Error trade-off

Fearnhead and Prangle 2012

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

1. Choice of summary:

$$\pi(\theta|D) \overset{?}{\approx} \pi(\theta|s_{obs})$$

2. Use of ABC acceptance kernel:

$$\pi(\theta|s_{obs}) \overset{?}{\approx} \pi_{ABC}(\theta|s_{obs})$$

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

Optimal (in some sense) to choose $\dim(s) = \dim(\theta)$

Machine learning invasion

ML algorithms are good at classification, usually better than humans.

ABC can be done via classification, albeit at the cost of abandoning the Bayesian interpretation.

Machine learning invasion

ML algorithms are good at classification, usually better than humans.

ABC can be done via classification, albeit at the cost of abandoning the Bayesian interpretation.
E.g. 1) Pudlo *et al.* 2015 and Marin *et al.* 2016 used random forests, others have used (C)NNs etc

1. Train a ML model, $m(X)$, to predict θ from D using a large number of simulator runs $\{\theta_i, X_i\}$
2. ABC then simulates θ from the prior and X from the simulator, and accepts θ if $m(X) \approx m(D_{obs})$

Machine learning invasion

ML algorithms are good at classification, usually better than humans.

ABC can be done via classification, albeit at the cost of abandoning the Bayesian interpretation.

E.g. 1) Pudlo *et al.* 2015 and Marin *et al.* 2016 used random forests, others have used (C)NNs etc

1. Train a ML model, $m(X)$, to predict θ from D using a large number of simulator runs $\{\theta_i, X_i\}$
2. ABC then simulates θ from the prior and X from the simulator, and accepts θ if $m(X) \approx m(D_{obs})$

E.g. 2) Generative Adversarial Networks (GANs, Goodfellow 2014) play a game between a generator and a discriminative classifier. The classifier tries to distinguish between data and simulation, and the generator tries to trick the classifier.

Machine learning invasion

ML algorithms are good at classification, usually better than humans.

ABC can be done via classification, albeit at the cost of abandoning the Bayesian interpretation.
E.g. 1) Pudlo *et al.* 2015 and Marin *et al.* 2016 used random forests, others have used (C)NNs etc

1. Train a ML model, $m(X)$, to predict θ from D using a large number of simulator runs $\{\theta_i, X_i\}$
2. ABC then simulates θ from the prior and X from the simulator, and accepts θ if $m(X) \approx m(D_{obs})$

E.g. 2) Generative Adversarial Networks (GANs, Goodfellow 2014) play a game between a generator and a discriminative classifier. The classifier tries to distinguish between data and simulation, and the generator tries to trick the classifier.
E.g. 3) Park *et al.* 2016, . . ., suggested using MMD in place of a vector of summaries, avoiding summarization.

Machine learning invasion

ML algorithms are good at classification, usually better than humans.

ABC can be done via classification, albeit at the cost of abandoning the Bayesian interpretation.

E.g. 1) Pudlo *et al.* 2015 and Marin *et al.* 2016 used random forests, others have used (C)NNs etc

1. Train a ML model, $m(X)$, to predict θ from D using a large number of simulator runs $\{\theta_i, X_i\}$
2. ABC then simulates θ from the prior and X from the simulator, and accepts θ if $m(X) \approx m(D_{obs})$

E.g. 2) Generative Adversarial Networks (GANs, Goodfellow 2014) play a game between a generator and a discriminative classifier. The classifier tries to distinguish between data and simulation, and the generator tries to trick the classifier.

E.g. 3) Park *et al.* 2016, . . ., suggested using MMD in place of a vector of summaries, avoiding summarization.

All work well in simulation studies where the model is well specified and there is a true θ...

Accelerating ABC with surrogates

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.

Surrogate ABC

- Wilkinson 2014
- Meeds and Welling 2014
- Gutmann and Corander 2015
- Strathmann, Sejdinovic, Livingstone, Szabo, Gretton 2015
- $\vdots$

With obvious influence from emulator community (e.g. Sacks, Welch, Mitchell, and Wynn 1989, Kennedy and O'Hagan 2001)

Constituent elements:

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

$\exists$ a relationship to probabilistic numerics

Target of approximation for the surrogate

- Simulator output within synthetic likelihood (Meeds et al 2014) e.g.

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

- (ABC) Likelihood type function (W. 2014)

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

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

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

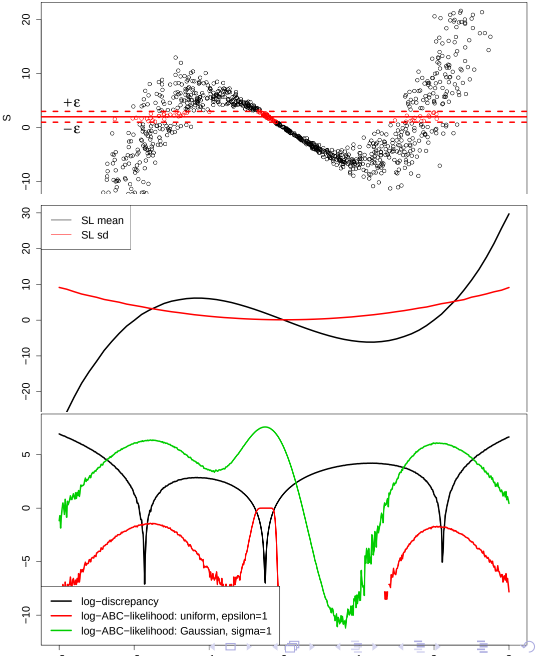
- Gradients (Strathmann et al 2015)

The difficulty of each approach depends on smoothness, dimension, focus etc.

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

Synthetic likelihood:

ABC likelihood and discrepancy:



Inference

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage - under-utilizes the surrogate, sacrificing speed-up.

Inference

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage - under-utilizes the surrogate, sacrificing speed-up.

Instead, Conrad *et al.* 2015 developed an intermediate approach that asymptotically samples from the exact posterior.

- proposes new θ - if uncertainty in surrogate prediction is such that it is unclear whether to accept or reject, then rerun simulator, else trust surrogate.

Inference

- Directly use the surrogate to calculate the posterior (Kennedy and O'Hagan 2001 etc) - over-utilizes the surrogate, sacrificing exact sampling.
- Correct for the use of a surrogate, e.g., using a Metropolis step (Rasmussen 2003, Sherlock *et al.* 2015, etc), which requires simulator evaluations at every stage - under-utilizes the surrogate, sacrificing speed-up.

Instead, Conrad *et al.* 2015 developed an intermediate approach that asymptotically samples from the exact posterior.

- proposes new θ - if uncertainty in surrogate prediction is such that it is unclear whether to accept or reject, then rerun simulator, else trust surrogate.

> *It is inappropriate to be concerned about mice when there are tigers abroad (Box 1976)*

Model discrepancy, ABC approximations, sampling errors etc may mean it is not worth worrying...

Acquisition rules

The key determinant of emulator accuracy is the design used to train the GP

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

Usual design choices are space-filling designs

- Maximin latin hypercubes, Sobol sequences

Acquisition rules

The key determinant of emulator accuracy is the design used to train the GP

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

Usual design choices are space-filling designs

- Maximin latin hypercubes, Sobol sequences

Calibration doesn't need a global approximation to the simulator - this is wasteful.

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

Cf David's talk

History matching waves

Craig *et al.* 1997

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

History matching waves

Craig *et al.* 1997

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

- But we only need to make good predictions near $\hat{\theta}$
- Introduce waves of history matching.
- In each wave, build a GP model that can rule out regions of space as implausible.

History matching waves

Craig *et al.* 1997

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

- But we only need to make good predictions near $\hat{\theta}$
- Introduce waves of history matching.
- In each wave, build a GP model that can rule out regions of space as implausible.

We decide that θ is implausible if

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

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

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

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

Choice of T is problem specific; start conservatively with T large and decrease

Example: Ricker Model

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.

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

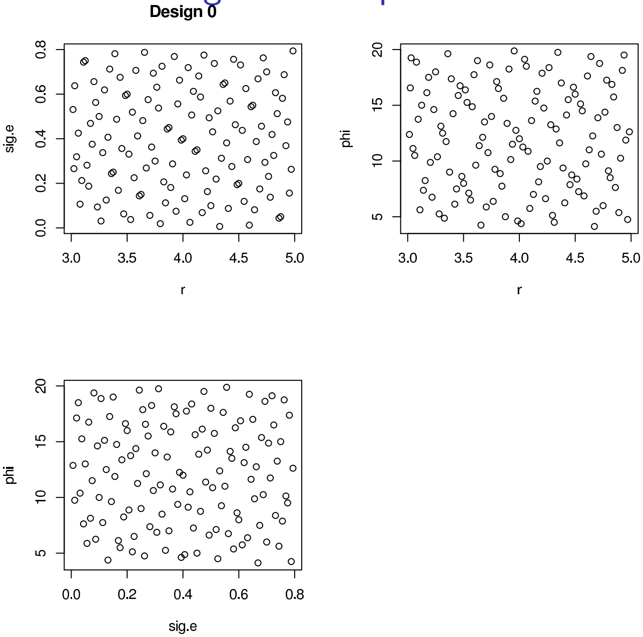
  where e_t are independent $N(0, \sigma_e^2)$ process noise

- Assume we observe counts y_t where

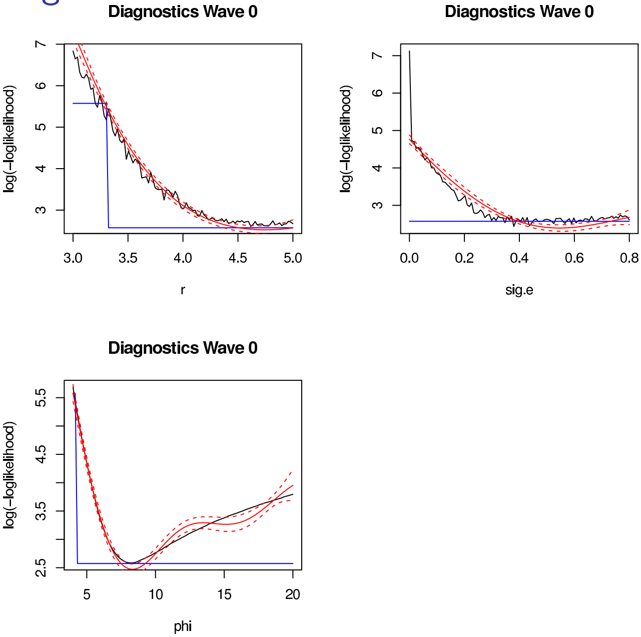
$$y_t \sim Po(\phi N_t)$$

Used in Wood to demonstrate the synthetic likelihood approach.

Results - Design 1 - 128 pts



Diagnostics for GP 1 - threshold = 5.6

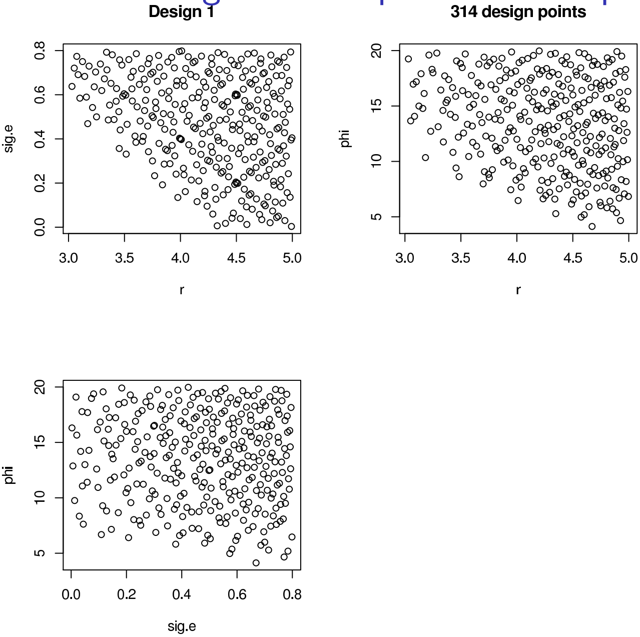


Diagnostics Wave 0

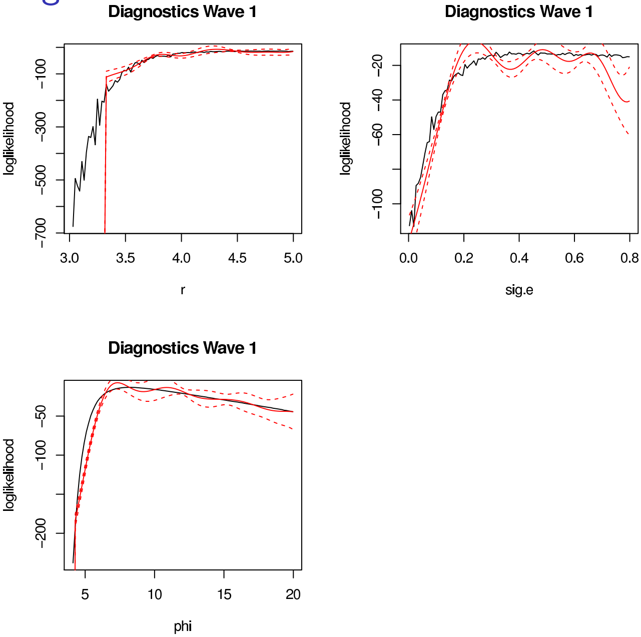
Diagnostics Wave 0

Diagnostics Wave 0

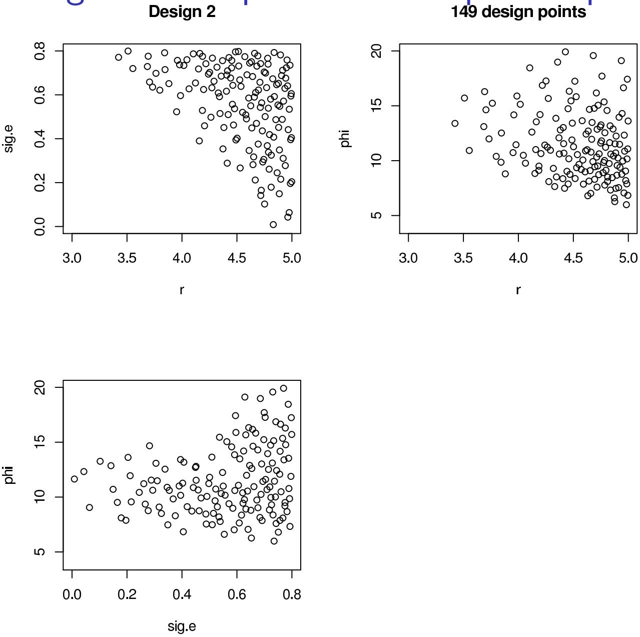
Results - Design 2 - 314 pts - 38% of space implausible



Diagnostics for GP 2 - threshold = -21.8



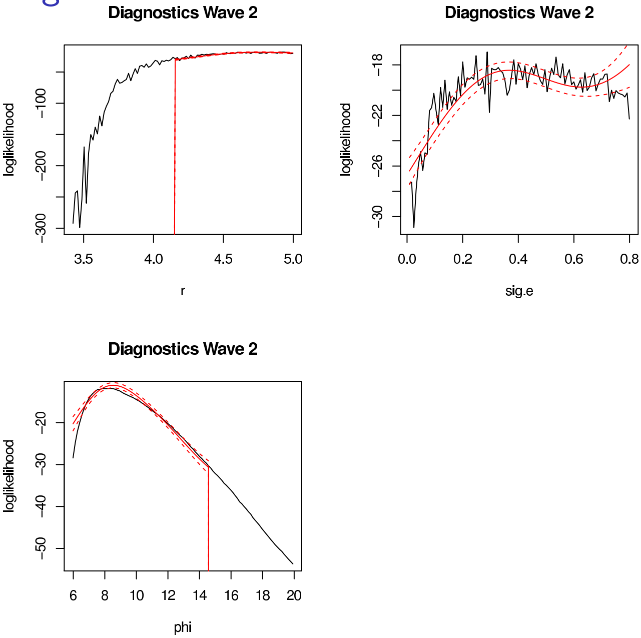
Design 3 - 149 pts - 62% of space implausible



Design 2

149 design points

Diagnostics for GP 3 - threshold = -20.7

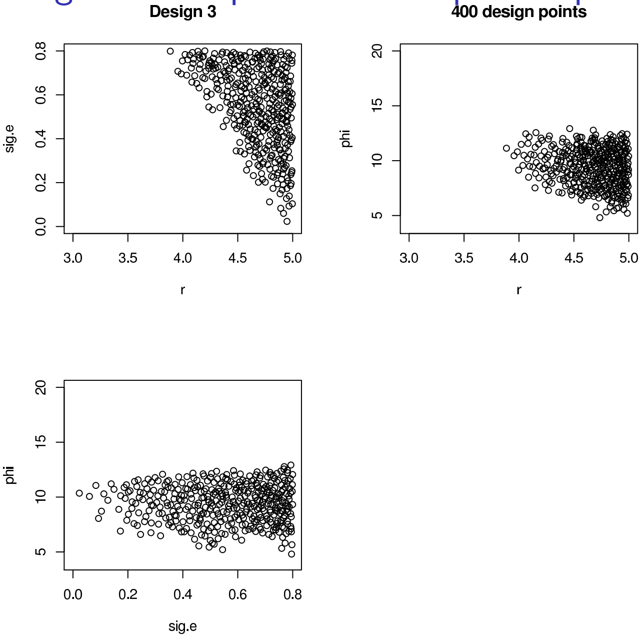


Diagnostics Wave 2

Diagnostics Wave 2

Diagnostics Wave 2

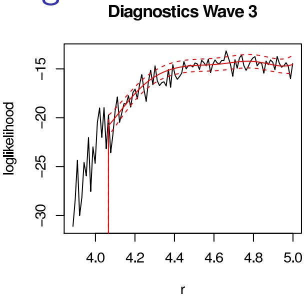
Design 4 - 400 pts - 95% of space implausible

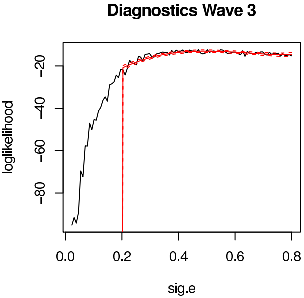


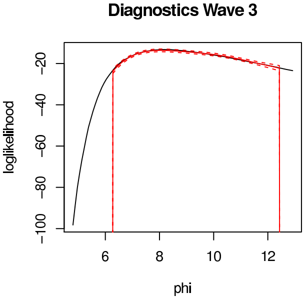
Design 3

400 design points

Diagnostics for GP 4 - threshold = -16.4

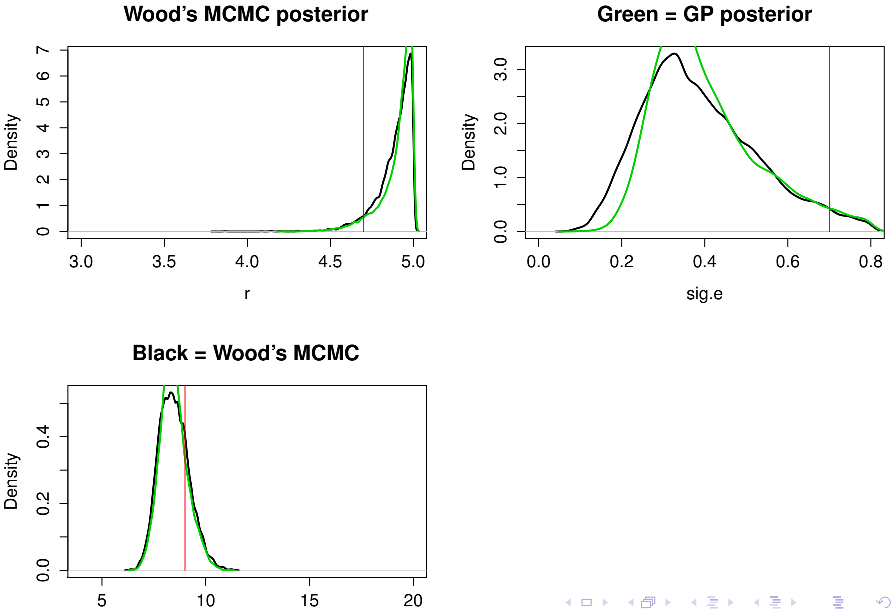






MCMC Results

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.

Inference for misspecified models

An appealing idea

Kennedy an O'Hagan 2001

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If $f_\theta(x)$ is our simulator, y the observation, then perhaps we can correct f by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta \sim GP$$

An appealing idea

Kennedy an O'Hagan 2001

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If $f_\theta(x)$ is our simulator, y the observation, then perhaps we can correct f by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta \sim GP$$

This greatly expands $\mathcal{F}$ into a non-parametric world.

An appealing, but flawed, idea

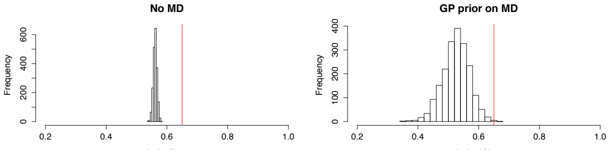
Kennedy and O'Hagan 2001, Brynjarsdottir and O'Hagan 2014

Simulator

$$f_\theta(x) = \theta x$$

Reality

$$g(x) = \frac{\theta x}{1 + \frac{x}{a}} \quad \theta = 0.65, a = 20$$



Bolting on a GP can correct your predictions, but won't necessarily fix your inference.

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is to complex to understand

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is to complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.

Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

    ie We never forget the prior, but the prior is to complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.
  - Wong et al 2017 impose identifiability (for δ and θ) by giving up and identifying

$$\theta^* = \arg\min_\theta \int (\zeta(x) - f_\theta(x))^2 d\pi(x)$$

History matching

ABC was proposed as a method of last resort, but there is evidence it works particularly well for mis-specified models.

History matching

ABC was proposed as a method of last resort, but there is evidence it works particularly well for mis-specified models.

History matching was designed for inference in mis-specified models. It seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\mathrm{ar}_{F_\theta}(Y)}}$$

History matching

ABC was proposed as a method of last resort, but there is evidence it works particularly well for mis-specified models.

History matching was designed for inference in mis-specified models. It seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\mathrm{ar}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta)\mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta, y) \leq \epsilon})$$

for some choice of S (typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$) and ϵ.

History matching

ABC was proposed as a method of last resort, but there is evidence it works particularly well for mis-specified models.

History matching was designed for inference in mis-specified models. It seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\mathbb{V}\mathrm{ar}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta)\mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta, y) \leq \epsilon})$$

for some choice of S (typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$) and ϵ.

They have thresholding of a score in common and are algorithmically comparable (thresholding).

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative
  - Bayes/Max-likelihood estimates usually concentrate asymptotically. If $G \notin \mathcal{F}$ can we hope to learn precisely about θ?

History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative
  - Bayes/Max-likelihood estimates usually concentrate asymptotically. If $G \notin \mathcal{F}$ can we hope to learn precisely about θ?

Conclusions

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

- not a silver bullet - if likelihood methods possible, use them instead (unless you are misspecified...)

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.

Machine learning approaches are now the largest area of research activity in ABC

Conclusions

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

- not a silver bullet - if likelihood methods possible, use them instead (unless you are misspecified...)

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.

Machine learning approaches are now the largest area of research activity in ABC

Thank you for listening!

r.d.wilkinson@sheffield.ac.uk

References - basics

Included in order of appearance in tutorial, rather than importance! Far from exhaustive - apologies to those I've missed (e.g. all those since 2014)

- 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.