GP-ABC: accelerating inference for intractable stochastic computer models

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

April 9, 2015
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?
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 $\theta$ and i.c.s and the simulator, $f(\theta)$, generates output $X$.
- The inverse-problem: observe data $D$, estimate parameter values $\theta$ 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 \( c_\theta \) 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 \( \theta \) 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. Accelerating ABC using meta-models
Basics
‘Likelihood-Free’ Inference

Rejection Algorithm

- Draw $\theta$ from prior $\pi(\cdot)$
- Accept $\theta$ with probability $\pi(D | \theta)$

Accepted $\theta$ are independent draws from the posterior distribution, $\pi(\theta | D)$. 
‘Likelihood-Free’ Inference

**Rejection Algorithm**

- Draw $\theta$ from prior $\pi(\cdot)$
- Accept $\theta$ with probability $\pi(D | \theta)$

Accepted $\theta$ are independent draws from the posterior distribution, $\pi(\theta | D)$.

If the likelihood, $\pi(D|\theta)$, is unknown:

**‘Mechanical’ Rejection Algorithm**

- Draw $\theta$ from $\pi(\cdot)$
- Simulate $X \sim f(\theta)$ from the computer model
- Accept $\theta$ if $D = X$, i.e., if computer output equals observation

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

If $\mathbb{P}(D)$ is small (or $D$ continuous), we will rarely accept any $\theta$. Instead, there is an approximate version:

**Uniform Rejection Algorithm**

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

$\epsilon$ 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 $\pi(\theta | D)$. 
Rejection ABC

If \( \mathbb{P}(D) \) is small (or \( D \) continuous), we will rarely accept any \( \theta \). Instead, there is an approximate version:

**Uniform Rejection Algorithm**

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

\( \epsilon \) 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 \( \pi(\theta | D) \).
\( \epsilon = 10 \)

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

\[ \rho(D, X) = |D - X|, \quad D = 2 \]
$\epsilon = 7.5$
\( \epsilon = 5 \)
$\epsilon = 2.5$
$\epsilon = 1$
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 $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ 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 $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ 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
W. 2008/2013

We wanted to solve the inverse problem

\[ D = f(\theta) \]

but instead ABC solves

\[ D = f(\theta) + e. \]
ABC as a probability model

W. 2008/2013

We wanted to solve the inverse problem

\[ D = f(\theta) \]

but instead ABC solves

\[ D = f(\theta) + e. \]

ABC gives ‘exact’ inference under a different model!

We can show that

**Proposition**

If \( \rho(D, X) = |D - X| \), then ABC samples from the posterior distribution of \( \theta \) 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 $(\theta, X)$ 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 $\theta$ if $\rho(D, X) \leq \epsilon$
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 \( (\theta, X) \) 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 \( \theta \) 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.

- We don’t need to assume uniform error!
Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance $\epsilon$ - controls the ‘ABC error’

- Summary statistic $S(D)$ - controls ‘information loss’
Key challenges for ABC

Accuracy in ABC is determined by

- **Tolerance $\epsilon$** - controls the ‘ABC error’
  - how do we find efficient algorithms that allow us to use small $\epsilon$ 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 $\epsilon$** - controls the ‘ABC error’
  - how do we find efficient algorithms that allow us to use small $\epsilon$ 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 $\epsilon$** - controls the ‘ABC error’
  - how do we find efficient algorithms that allow us to use small $\epsilon$ 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 $\epsilon$, 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
We are targeting the joint distribution

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

To explore the $$(\theta, x)$$ 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

$$r = \frac{\pi_{ABC}(\theta', x'|D)Q((\theta', x'), (\theta, x))}{\pi_{ABC}(\theta, x|D)Q((\theta, x), (\theta', x'))}$$
We are targeting the joint distribution

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

To explore the $(\theta, x)$ 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

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

$$= \frac{\pi_\epsilon(D|x')\pi(x'|\theta')\pi(\theta')q(\theta', \theta)\pi(x|\theta)}{\pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)q(\theta, \theta')\pi(x'|\theta')}$$
We are targeting the joint distribution

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

To explore the \((\theta, x)\) 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

\[
r = \frac{\pi_{ABC}(\theta', x'|D)Q((\theta', x'), (\theta, x))}{\pi_{ABC}(\theta, x|D)Q((\theta, x), (\theta', x'))}
= \frac{\pi_\epsilon(D|x')\pi(x'|\theta')\pi(\theta')q(\theta', \theta)\pi(x|\theta)}{\pi_\epsilon(D|x)\pi(x|\theta)\pi(\theta)q(\theta, \theta')\pi(x'|\theta')}
= \frac{\pi_\epsilon(D|x')q(\theta', \theta)\pi(\theta')}{\pi_\epsilon(D|x)q(\theta, \theta')\pi(\theta)}
\]
This gives the following MCMC algorithm

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

1. Propose a move from $z_t = (\theta, x)$ to $(\theta', x')$ 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 $\epsilon$ is small.

Lee 2012 introduced several alternative MCMC kernels that are variance bounding and geometrically ergodic.
This gives the following MCMC algorithm

**MH-ABC - \( P_{Mar} (\theta_0, \cdot) \)**

1. Propose a move from \( z_t = (\theta, x) \) to \( (\theta', x') \) 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 \( \epsilon \) is small.

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

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 $\pi_t$ be the ABC distribution found by the ABC algorithm when we use tolerance $\epsilon_t$. 
Specifically, define a sequence of target distributions

\[
\pi_t(\theta, x) = \frac{\prod_{\rho(D,x) < \epsilon_t} \pi(x|\theta) \pi(\theta)}{C_t} = \frac{\gamma_t(\theta, x)}{C_t}
\]
At each stage \( t \), we aim to construct a weighted sample of particles that approximates \( \pi_t(\theta, x) \).

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

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 \( \epsilon \) is our simulator discrepancy

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

\[ D = \zeta + e \]

3. Declare \( \theta \) implausible if, e.g.,

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

where \( \sigma^2 \) is the combined variance implied by the emulator, discrepancy and measurement error.
History-matching

If $\theta$ 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 (NROY).

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 $\mu_\theta$ and covariance $\Sigma_\theta$ of the summary of $f(\theta)$ run $n$ times at $\theta$, 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
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.
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

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

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 $\theta$ 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
     \[
     m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot, \cdot)) \quad \text{and} \quad \nu(\theta) = \mathbb{V}ar f(\theta) \sim GP(0, c(\cdot, \cdot))
     \]
     and then assume
     \[
     f(\theta) \sim N(m(\theta), \nu(\theta))
     \]
     Cf. Wood 2010 synthetic likelihood approach.
**Emulating stochastic models**

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

1. **Model summaries of the simulator response:**
   - e.g., model
     \[
     m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot, \cdot)) \text{ and } v(\theta) = \mathbb{V}ar f(\theta) \sim GP(0, c(\cdot, \cdot))
     \]
     and then assume
     \[
     f(\theta) \sim N(m(\theta), v(\theta))
     \]
     Cf. Wood 2010 synthetic likelihood approach.

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
     \[
     m(\theta) = \mathbb{E}f(\theta) \sim GP(0, c(\cdot, \cdot)) \quad \text{and} \quad v(\theta) = \text{Var}f(\theta) \sim GP(0, c(\cdot, \cdot))
     \]
   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 \( \theta \).
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 \( \theta \)
- 1d response surface
- **But**, it can be hard to model.
The GABC framework assumes

\[ \pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)dX \]

\[ \approx \frac{1}{N} \sum \pi(D|X_i) \]

where \( X_i \sim \pi(X|\theta) \).
Likelihood estimation

The GABC framework assumes

\[ \pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)dX \]

\[ \approx \frac{1}{N} \sum \pi(D|X_i) \]

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

\[ l(\theta) = \log L(\theta) \]
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

W. 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.
- In each wave, build a GP model that can rule out regions of space as implausible.
Implausibility

Given a model of the likelihood

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

we define the plausible/NROY set

\[ \mathcal{P}_\theta = \{ \theta \in \Theta : l(\theta) \geq T \} \]
Implausibility

Given a model of the likelihood

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

we define the plausible/NROY set

\[ \mathcal{P}_\theta = \{ \theta \in \Theta : l(\theta) \geq T \} \]

- 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 \( \theta \) 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-implausible range and build a new Gaussian process.

This new GP will lead to a new definition of implausibility.

\[ \ldots \]
• 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-implausible range and build a new Gaussian process

• This new GP will lead to a new definition of implausibility

• …

This is essentially a classification problem. We use a conservative decision rule, declaring \( \theta \) implausible if

\[
m(\theta) + 3\sigma < T
\]

Equivalently,

\[
P(l(\theta) < T) > 0.997
\]
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 $t$.

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

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

Design 0

- Scatter plots showing the relationship between parameters (e.g., $\sigma_e$, $\phi$, $r$) and their ranges.
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

Diagnostics Wave 1

loglikelihood vs. $r$

loglikelihood vs. $\text{sig.e}$

Diagnostics Wave 1

loglikelihood vs. $\phi$
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

Wood’s MCMC posterior

Green = GP posterior

Black = Wood’s MCMC
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.
Active learning for history-matching/GP-ABC

Work with James Hensman

Using a space filling design is often not optimal. Instead, we can use active learning ideas from machine learning/Bayesian optimisation/probabilistic numerics to sequentially build a design $\theta_1, \theta_2, \ldots$

One option is to minimise the expected entropy of the resulting history match....
Active learning for history-matching/GP-ABC

Work with James Hensman

Using a space filling design is often not optimal. Instead, we can use active learning ideas from machine learning/Bayesian optimisation/probabilistic numerics to sequentially build a design $\theta_1, \theta_2, \ldots$

One option is to minimise the expected entropy of the resulting history match.....

- Any GP emulator allows us to calculate a probabilistic classification
  
  \[ p(\theta) = \mathbb{P}(\theta \in \mathcal{P}_\theta) \]

- The entropy of our belief at $\theta$ is

  \[ E(\theta) = -p \log p - (1 - p) \log(1 - p) \]

- We could choose the next design point, $\theta_{n+1}$, to minimize $E(\theta)$, a point wise criteria.

- This is numerically simple, but the additional design points tend to accumulate on the edge of the domain $\Theta$. 
Instead, we can find the average entropy

$$E_n = \int E(\theta) d\theta$$

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

- Choose the next design point to minimise the expected average entropy

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

where

$$J_n(\theta) = \mathbb{E}(E_{n+1}|\theta_{n+1} = \theta)$$
EPm: toy climate model

- $DT_{crit\_conv}$ - critical temperature gradient that triggers convection
- GAMMA - emissivity parameter for water vapour
- Calibrate to global average surface temperature
Finding \( \theta \) which minimises \( J_n(\theta) = \mathbb{E}(E_{n+1}|\theta_{n+1} = \theta) \) is expensive.

We can use Bayesian optimization to find the optima:

1. Evaluate \( J_n(\theta) \) at a small number of locations
2. Build a GP model of \( J_n(\cdot) \)
3. Choose the next \( \theta \) at which to evaluate \( J_n \) so as to minimise the EI criterion
4. Return to step 2.
History match

Can we learn the following plausible set?

- A sample from a GP on $\mathbb{R}^2$.
- Find $x$ s.t. $-2 < f(x) < 0$
Iterations 10 and 15
Left = p(θ), middle = E(θ), right = \tilde{J}(θ)
Iterations 15

Left = $p(\theta)$, middle = $E(\theta)$, right = $\tilde{J}(\theta)$
Iterations 20 and 24

Video
Questions, challenges, difficulties

- The use of waves is unsatisfactory. Would prefer a global approximation, perhaps using logistic regression:

$$n_{acc} \sim \text{Bin}(N_{\text{trial}}, p(\theta)) \quad \text{logit } p(\cdot) \sim GP(m(\cdot), c(\cdot, \cdot))$$

- Currently, each wave considered in turn. Classification errors in earlier waves can never be corrected. Is it possible to use a design criterion that operates across waves?

- Classification rule

$$\theta \text{ implausible if } l(\theta) < T$$

is entirely heuristic, but theory should be possible....

- Efficient calculation of the posterior given the final emulator: HMC-NUTS, ...

::
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.

Thank you for listening!
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.

Thank you for listening!

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

- Wilkinson, *JMLR*, 2013
- Chevalier, Picheny, Ginsbourger, Comp. Stat. and Data Anal., 2014