Part I: **Approximate** inference for approximate computer models

Richard Wilkinson
axtis21

School of Mathematical Sciences
University of Nottingham

September 6, 2021

Talk plan

Today - tutorial

- Approximate Bayesian computation (ABC) for complex models
- Accelerating ABC via sampling and summaries
- Surrogate models for ABC

Tomorrow - speculation

- Inference for misspecified models
- Variational inference and generalizations

Bayesian inverse problems

Calibration, tuning, model fitting, parameter estimation, inference....

Components:

- Simulator (mechanistic model) f that takes unknown parameters θ as an input (as well as ICs, control variables etc.) and generates output X

$$X = f(\theta).$$

May be stochastic ($f(\theta) = f(\theta, U)$) or deterministic.

Bayesian inverse problems

Calibration, tuning, model fitting, parameter estimation, inference....

Components:

- Simulator (mechanistic model) f that takes unknown parameters θ as an input (as well as ICs, control variables etc.) and generates output X

$$X = f(\theta).$$

  May be stochastic ($f(\theta) = f(\theta, U)$) or deterministic.
- Statistical model that relates f to the observed data D, e.g.,

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

Bayesian inverse problems

Calibration, tuning, model fitting, parameter estimation, inference....

Components:

- Simulator (mechanistic model) f that takes unknown parameters θ as an input (as well as ICs, control variables etc.) and generates output X

$$X = f(\theta).$$

  May be stochastic ($f(\theta) = f(\theta, U)$) or deterministic.
- Statistical model that relates f to the observed data D, e.g.,

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

- The inverse-problem: find parameter values θ which are consistent with the data and the model

The Bayesian approach is to find the posterior distribution

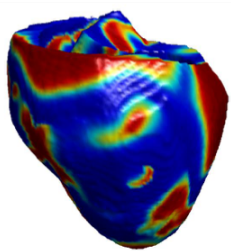
$$\pi(\theta|D) \propto \pi(\theta)\pi(D|\theta)$$

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

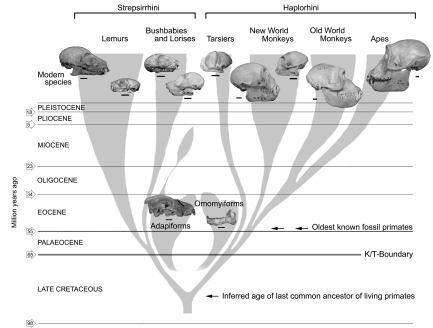
Examples

Atrial fibrillation

Simulation of electrical activation on the left atrium. Unknown tissue properties need to be estimated from noisy sparse ECG and MRI data. Estimates used to guide surgery.



Simulation of primate evolution, with unknown origination time to be estimated from fossil and genetic record.



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)
- 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)

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 widely used primarily because they are

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

Basics of ABC

‘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:

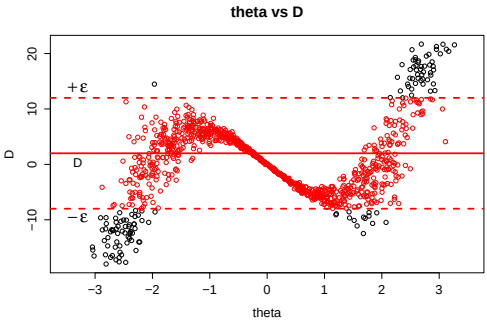
Uniform Rejection Algorithm

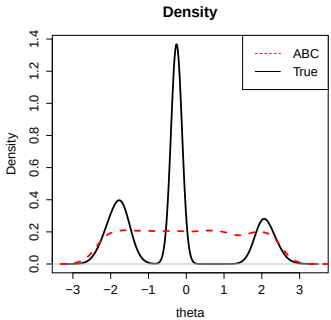
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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

$\epsilon = 10$

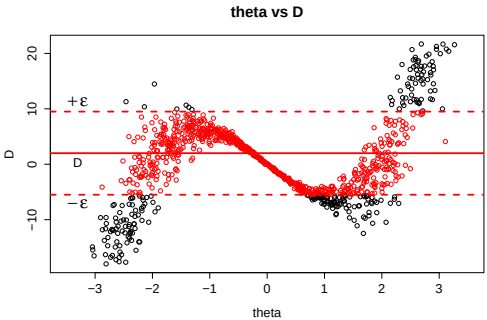


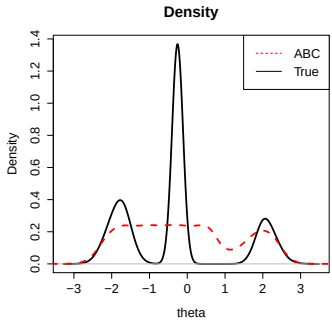


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 7.5$

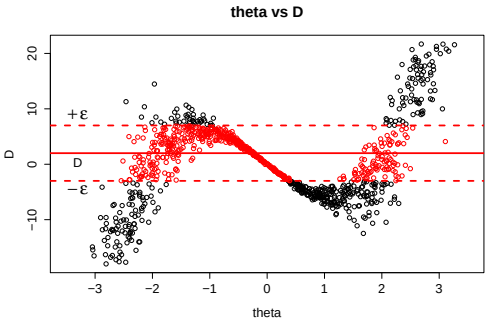


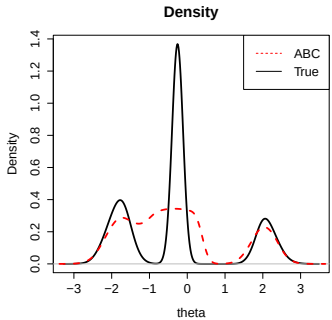


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 5$

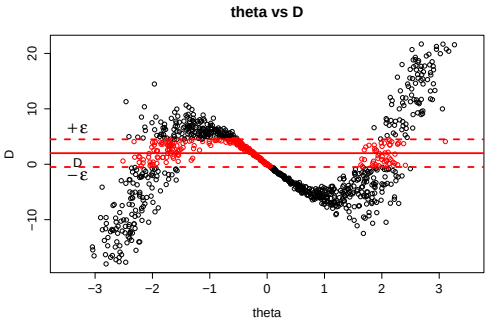


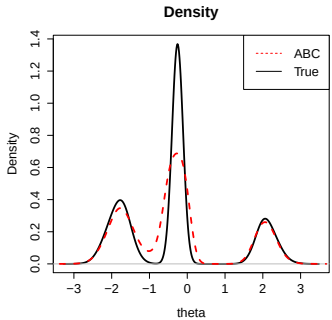


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 2.5$

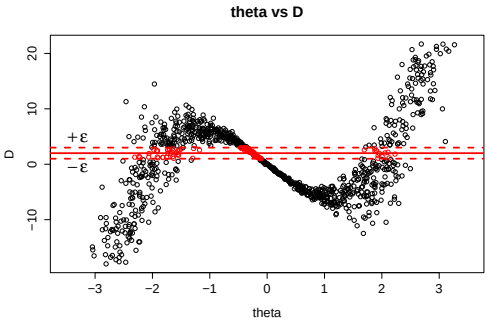


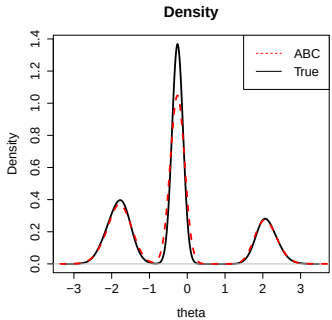


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 1$





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

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

Rejection ABC

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

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

Approximate Rejection Algorithm With Summaries

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

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

Rejection ABC

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

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

Approximate Rejection Algorithm With Summaries

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

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

Simple $\rightarrow$ Popular with non-statisticians

Key challenges for ABC

Scoring θ

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

Computing acceptable θ

- 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 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, W. 2013 ...

We are targeting the joint distribution

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

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012, W. 2013 ...

We are targeting the joint distribution

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

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. 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, W. 2013 ...

We are targeting the joint distribution

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

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. The Metropolis-Hastings (MH) acceptance probability is then

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

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012, W. 2013 ...

We are targeting the joint distribution

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

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. The Metropolis-Hastings (MH) acceptance probability is then

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

MCMC-ABC

Marjoram *et al.* 2003, Sisson and Fan 2011, Lee 2012, W. 2013 ...

We are targeting the joint distribution

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

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. The Metropolis-Hastings (MH) acceptance probability is then

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

NB: HMC is not possible (w/o a surrogate)

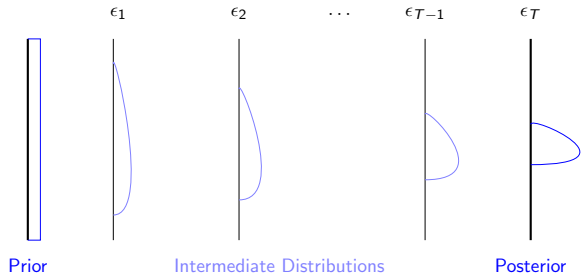
Sequential ABC algorithms

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

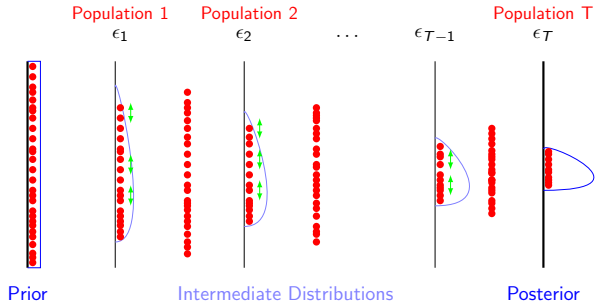
Choose a sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ and let π_t be the ABC approximation when using tolerance ϵ_t.

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 $\pi_t(\theta, x)$.



Picture from Toni and Stumpf 2010

Model selection

W. 2007, Grelaud *et al.* 2009

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

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

References:

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

SOMETHING ABOUT ML APPROACHES MMD and Neural Nets...

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

But if we know of a sufficient summary, then inference with $S(D)$ can be much quicker when $\dim S(D) \ll \dim D$.

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

But if we know of a sufficient summary, then inference with $S(D)$ can be much quicker when $\dim S(D) \ll \dim D$.

However, low-dimensional sufficient statistics are rarely available.
Instead, we focus on choosing low dimensional summaries that are good enough.

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

The first approximation allows the matching between $s_{obs} = S(D)$ 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:
$$\pi(\theta|D) \stackrel{?}{\approx} \pi(\theta|s_{obs})$$
2. Use of ABC acceptance kernel:
$$\pi(\theta|s_{obs}) \stackrel{?}{\approx} \pi_{ABC}(\theta|s_{obs})$$

The first approximation allows the matching between $s_{obs} = S(D)$ 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 approaches

Machine learning approaches

1. Automated summaries: Use random forests, NNs etc to generate a summary (see Raynal et al. 2019 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 approaches

1. Automated summaries: Use random forests, NNs etc to generate a summary (see Raynal et al. 2019 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 $m(X) \approx m(D_{obs})$
2. Generative Adversarial Networks (GANs, Mohamed et al. 2017 etc) 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 approaches

1. Automated summaries: Use random forests, NNs etc to generate a summary (see Raynal et al. 2019 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 $m(X) \approx m(D_{obs})$
2. Generative Adversarial Networks (GANs, Mohamed et al. 2017 etc) 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.
3. Kernel methods: e.g. use kernel mean embedding of the distribution (MMD) to avoid the need to summarize - inference is then *simply* projection in the RKHS.
4. A variety of neural network approaches to directly approximate the posterior, see e.g. normalizing flows (Papamakarios et al. 2021).

Machine learning approaches

1. Automated summaries: Use random forests, NNs etc to generate a summary (see Raynal et al. 2019 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 $m(X) \approx m(D_{obs})$
2. Generative Adversarial Networks (GANs, Mohamed et al. 2017 etc) 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.
3. Kernel methods: e.g. use kernel mean embedding of the distribution (MMD) to avoid the need to summarize - inference is then *simply* projection in the RKHS.
4. A variety of neural network approaches to directly approximate the posterior, see e.g. normalizing flows (Papamakarios et al. 2021).

NB: beware of all automated summary selection approaches if misspecified

Accelerating ABC with surrogates

References:

- W. 2014
- Meeds and Welling 2014
- Gutmann and Corander 2015
- Strathmann, Sejdinovic, Livingstone, Szabo, Gretton 2015
- ELFI and BOLFI software
- .
- :

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

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.

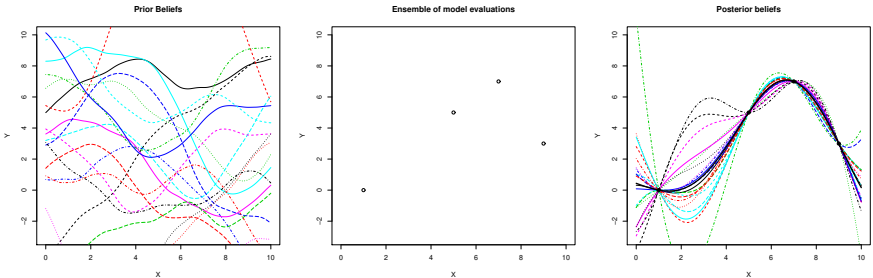
However,

- 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

If the simulator f is computationally expensive, we can build a surrogate/emulator $\tilde{f}$.



We can then perform inference with the emulator, accounting for the approximation error.

Constituent elements:

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

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)

$$L_{ABC}(\theta) = \mathbb{E}_{X|\theta} K_\epsilon[\rho(S(D), S(X))] \equiv \mathbb{E}_{X|\theta} \pi_\epsilon(D|X)$$
$$\approx \frac{1}{N} \sum_{i=1}^{N} \pi_\epsilon(D|X_i = f(\theta, U_i))$$

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

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

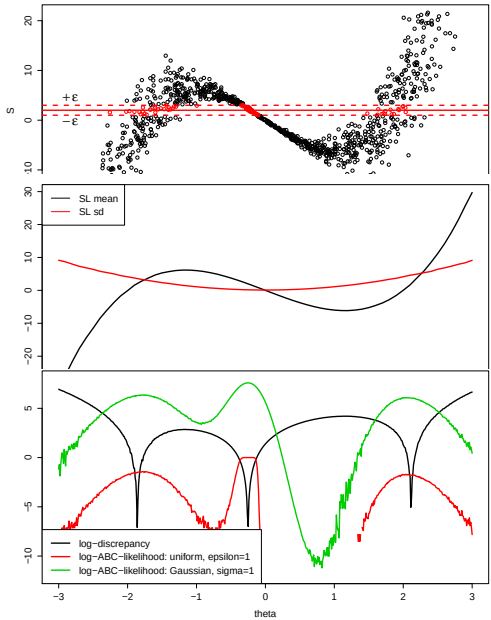
- Gradients (Strathmann et al 2015)

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 and others have developed intermediate approaches that asymptotically sample 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 and others have developed intermediate approaches that asymptotically sample 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.

Function approximation where it matters

W. 2014

The log-likelihood $l(\theta) = \log L(\theta)$ can vary over several orders of magnitude, and thus many surrogate models can struggle to accurately approximate it.

Function approximation where it matters

W. 2014

The log-likelihood $l(\theta) = \log L(\theta)$ can vary over several orders of magnitude, and thus many surrogate models can struggle to accurately approximate it.

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

Function approximation where it matters

W. 2014

The log-likelihood $l(\theta) = \log L(\theta)$ can vary over several orders of magnitude, and thus many surrogate models can struggle to accurately approximate it.

- But we only need good predictions near $\hat{\theta}$
- Introduce waves of history matching/ sequential batch design.
- 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 $\pi(\theta|y) = 0$

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.

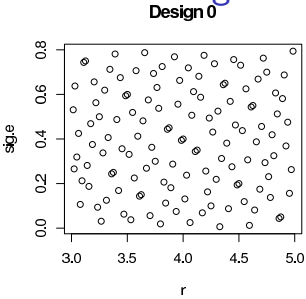
$$N_{t+1} = rN_t e^{-N_t + e_r}$$

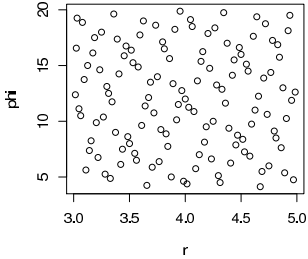
  where e_t are independent $N(0, \sigma_e^2)$ process noise
- Assume we observe counts y_t where

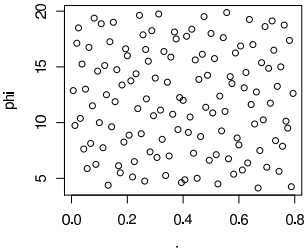
$$y_t \sim Po(\phi N_t)$$

Used in Wood to demonstrate the synthetic likelihood approach.

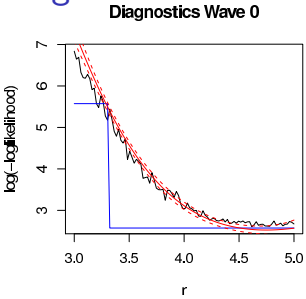
Results - Design 1 - 128 pts

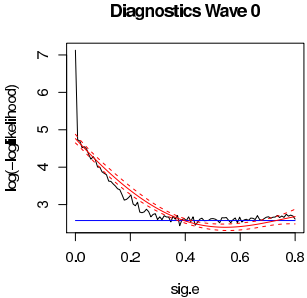


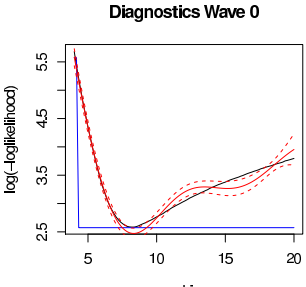




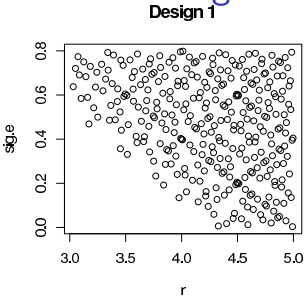
Diagnostics for GP 1 - threshold = 5.6

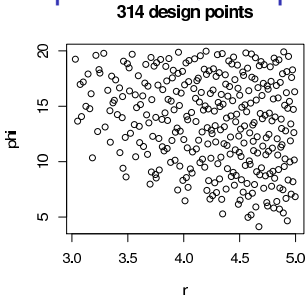


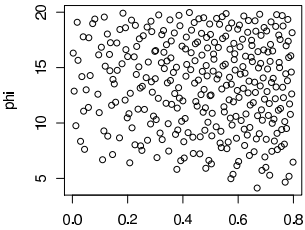




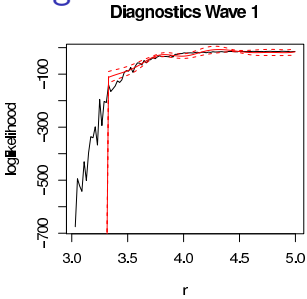
Results - Design 2 - 314 pts - 38% of space implausible

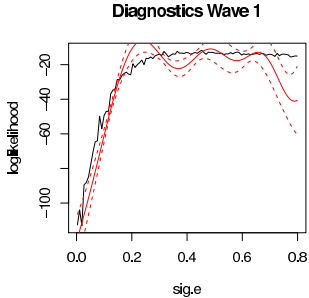


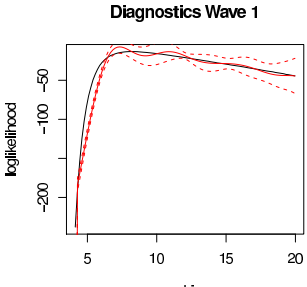




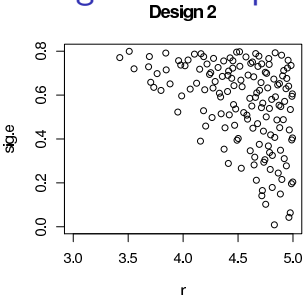
Diagnostics for GP 2 - threshold = -21.8

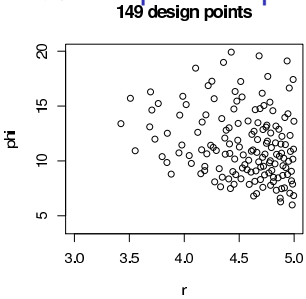


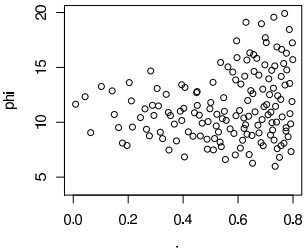




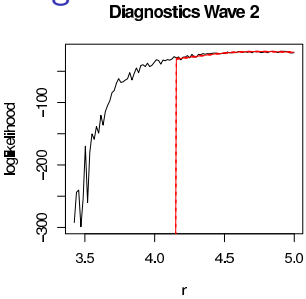
Design 3 - 149 pts - 62% of space implausible

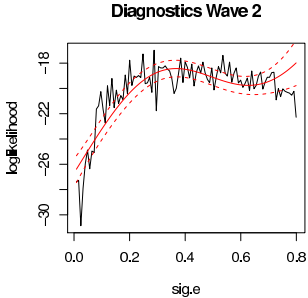


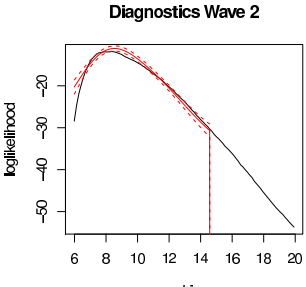




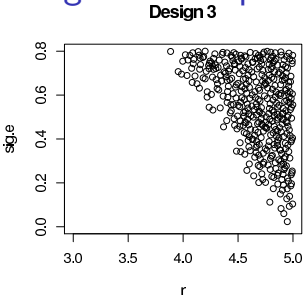
Diagnostics for GP 3 - threshold = -20.7

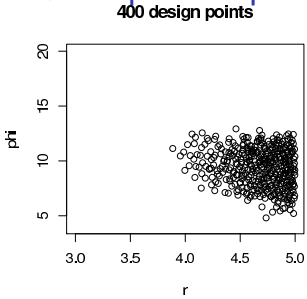


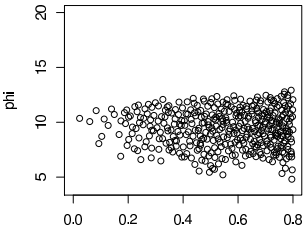




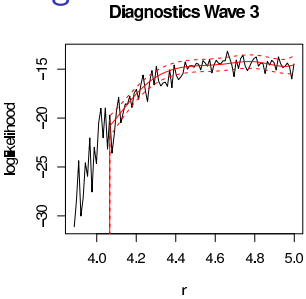
Design 4 - 400 pts - 95% of space implausible

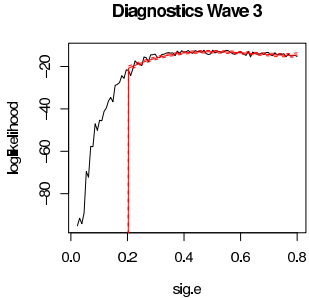


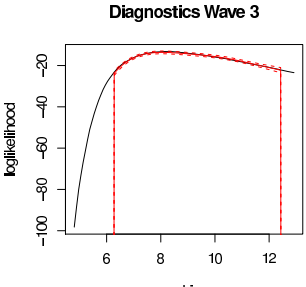




Diagnostics for GP 4 - threshold = -16.4

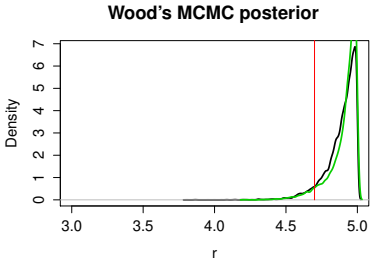


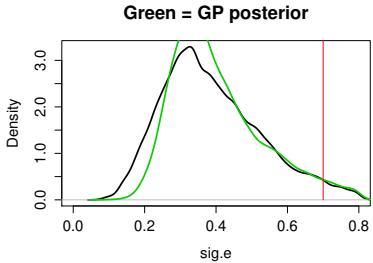


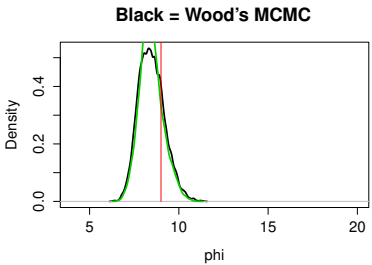


MCMC Results

Comparison with Wood 2010, synthetic likelihood approach







Computational details

- The Wood MCMC method used $10^5 \times 500$ simulator runs
- The GP code used $(128 + 314 + 149 + 400) = 991 \times 500$ simulator runs
  - 1/100th of the number used by Wood's method.

By the final iteration, the Gaussian processes had ruled out over 98% of the original input space as implausible,

- the MCMC sampler did not need to waste time exploring those regions.

Conclusions

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

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

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

- May need to go further and use surrogate models...

For misspecified models, focusing on doing approximate Bayesian inference with a small approximation error may not be a good use of resource.

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.

- May need to go further and use surrogate models...

For misspecified models, focusing on doing approximate Bayesian inference with a small approximation error may not be a good use of resource.

Thank you for listening!