Quiz

Suppose we are testing H_0 vs H_1 and find $p < 0.05$ and reject H_0 at the 5% level. Which, if any, of these statements is true:

- The probability H_0 is true is 0.05
- If H_0 is false, then we will correctly reject H_0 95% of the time.
- If we observe $p = 0.05$, the probability we falsely reject H_0 given that it is true is 0.05.
- If H_0 is true, we will falsely reject it 5% of the time

Quiz

Suppose we are testing H_0 vs H_1 and find $p < 0.05$ and reject H_0 at the 5% level.
Which, if any, of these statements is true:

- The probability H_0 is true is 0.05 False
- If H_0 is false, then we will correctly reject H_0 95% of the time. False
- If we observe $p = 0.05$, the probability we falsely reject H_0 given that it is true is 0.05.False
- If H_0 is true, we will falsely reject it 5% of the time True

Personal view: p-values are not the problem - it is the belief that every dataset should give a yes/no answer.

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

Richard Wilkinson
Darmstadt

School of Mathematics and Statistics
University of Sheffield

October 10, 2018

Why be Bayesian?

Why ~~be Bayesian~~ do Bayesian analyses?

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.
- Build in expert belief / rule out things we know are unlikely / regularise the solution

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.
- Build in expert belief / rule out things we know are unlikely / regularise the solution
- It is all just probability
  - makes combining different uncertainties easy/possible e.g. calibrated prediction $p(y|x) = \int p(y|\theta)p(\theta|x)\mathrm{d}\theta$
  - deals with equifinality (ie multiple feasible values, under-specified systems)
  - Simpler! Q:What's the difference between probability, significance, coverage, confidence, p-values etc

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.
- Build in expert belief / rule out things we know are unlikely / regularise the solution
- It is all just probability
  - makes combining different uncertainties easy/possible e.g. calibrated prediction $p(y|x) = \int p(y|\theta)p(\theta|x)\mathrm{d}\theta$
  - deals with equifinality (ie multiple feasible values, under-specified systems)
  - Simpler! Q:What's the difference between probability, significance, coverage, confidence, p-values etc
- It is possible / increasingly easy to do Bayesian inference
  - Frequentist procedures are 'mathematically challenging' to derive for complex models

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.
- Build in expert belief / rule out things we know are unlikely / regularise the solution
- It is all just probability
  - makes combining different uncertainties easy/possible e.g. calibrated prediction $p(y|x) = \int p(y|\theta)p(\theta|x)\mathrm{d}\theta$
  - deals with equifinality (ie multiple feasible values, under-specified systems)
  - Simpler! Q:What's the difference between probability, significance, coverage, confidence, p-values etc
- It is possible / increasingly easy to do Bayesian inference
  - Frequentist procedures are 'mathematically challenging' to derive for complex models

The downsides:

- We have to choose a prior.
  - In practice, 'priors of convenience' are often used. You need to really care about the answer to bother with expert elicitation.
  - You can check robustness wrt your choices.

Why ~~be Bayesian~~ do Bayesian analyses?

- Coherence: under various sets of axioms, Bayes is the only sensible choice.... cf Jaynes, de Finetti, Jeffreys etc.
- Build in expert belief / rule out things we know are unlikely / regularise the solution
- It is all just probability
  - makes combining different uncertainties easy/possible e.g. calibrated prediction $p(y|x) = \int p(y|\theta)p(\theta|x)\mathrm{d}\theta$
  - deals with equifinality (ie multiple feasible values, under-specified systems)
  - Simpler! Q:What's the difference between probability, significance, coverage, confidence, p-values etc
- It is possible / increasingly easy to do Bayesian inference
  - Frequentist procedures are 'mathematically challenging' to derive for complex models

The downsides:

- We have to choose a prior.
  - In practice, 'priors of convenience' are often used. You need to really care about the answer to bother with expert elicitation.
  - You can check robustness wrt your choices.
- There is no need for your posterior to relate to the world
  - Post-hoc checks (calibration etc) can help, but there are no frequency guarantees

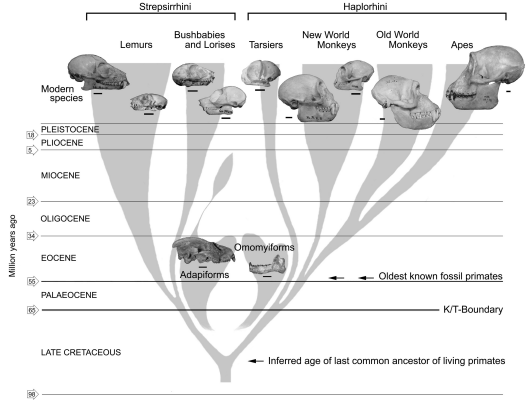
Calibration

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

The Bayesian approach is to find the posterior distribution

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

posterior $\propto$ prior $\times$ likelihood



Intractability

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

- usual intractability in Bayesian inference is not knowing $\pi(D)$.
- a problem is doubly intractable if $\pi(D|\theta) = c_\theta p(D|\theta)$ with c_θ unknown (cf Murray, Ghahramani and MacKay 2006)
- a problem is completely intractable if $\pi(D|\theta)$ is unknown and can't be evaluated (unknown is subjective). I.e., if the analytic distribution of the simulator, $f(\theta)$, run at θ is unknown.

Completely intractable models are where we need to resort to ABC methods

Approximate Bayesian Computation (ABC)

If the likelihood function is intractable, then ABC (approximate Bayesian computation) is one of the few approaches we can use to do inference.

Approximate Bayesian Computation (ABC)

If the likelihood function is intractable, then ABC (approximate Bayesian computation) is one of the few approaches we can use to do inference.

ABC algorithms are a collection of Monte Carlo methods used for calibrating simulators

- they do not require explicit knowledge of the likelihood function
- inference is done using simulation from the model (they are 'likelihood-free').

Approximate Bayesian Computation (ABC)

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

Plan

i. Basics
ii. Efficient sampling algorithms
iii. Regression adjustments/ post-hoc corrections
iv. Summary statistics
v. 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:

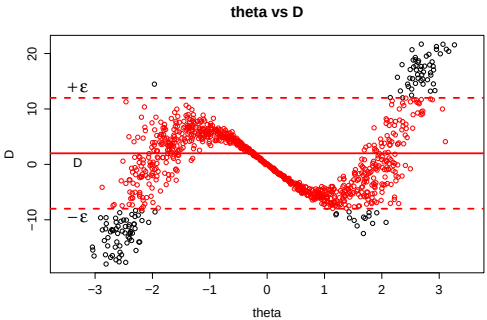
Uniform Rejection Algorithm

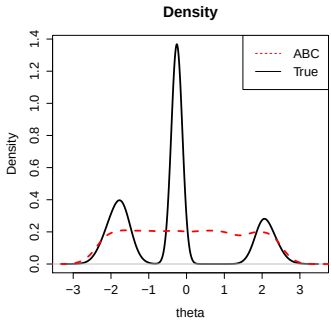
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(D, X) \leq \epsilon$

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

$\epsilon = 10$

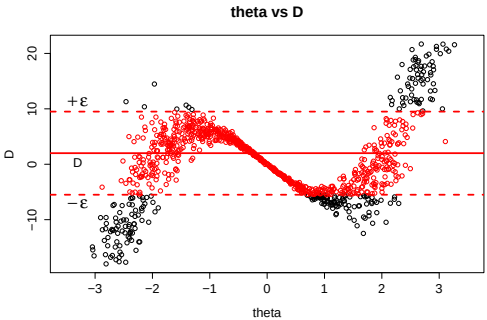


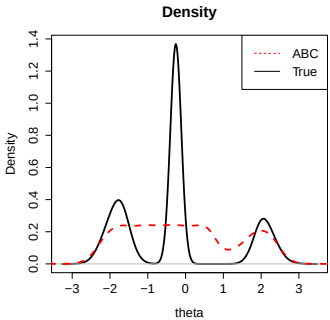


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 7.5$

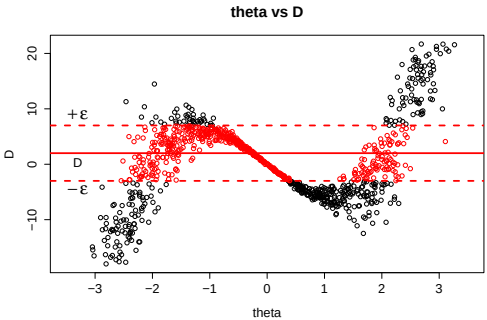


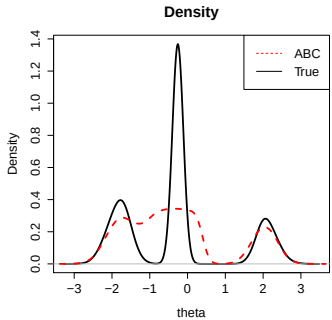


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 5$

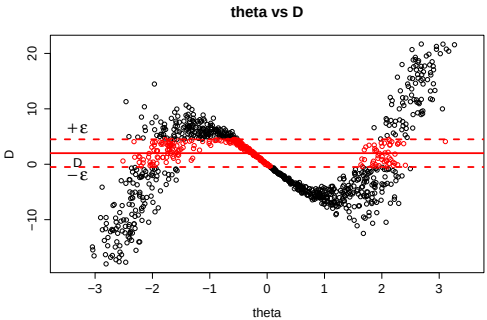


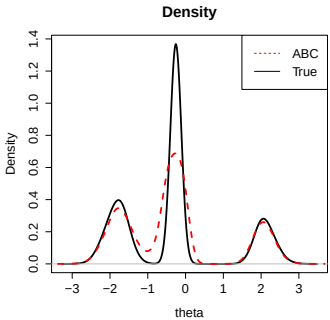


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 2.5$

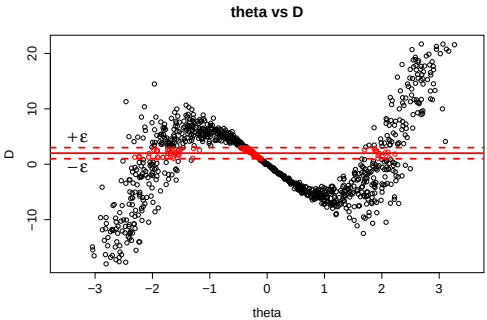


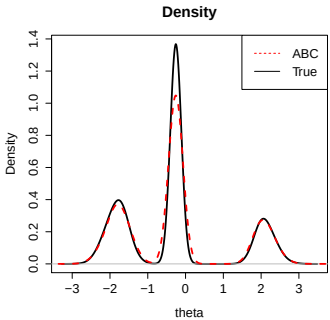


$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

$\epsilon = 1$

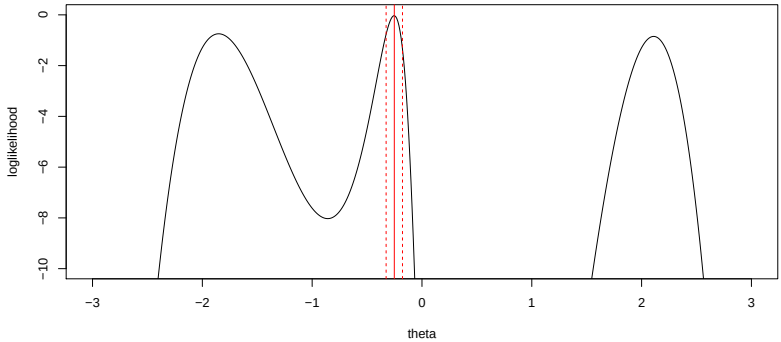




$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

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

Aside: Maximum likelihood and MCMC



$$\hat{\theta} = \arg\max L(\theta) = -0.252 \text{ and } \frac{d^2}{d\theta^2} \log L(\theta)\Big|_{\theta=\hat{\theta}} = -721$$

And so 'assuming' $\hat{\theta} \sim N(\theta, \mathcal{I}^{-1})$ gives a 95% confidence interval for θ of $[-0.33, -0.18]$

Aside: Maximum likelihood and MCMC

Does the difference matter if I only care about prediction?

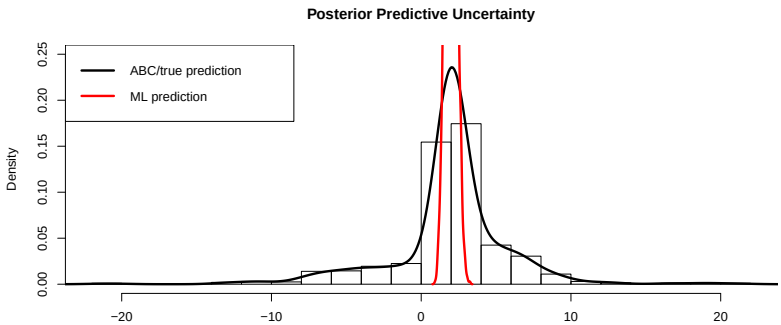
Aside: Maximum likelihood and MCMC

Does the difference matter if I only care about prediction?

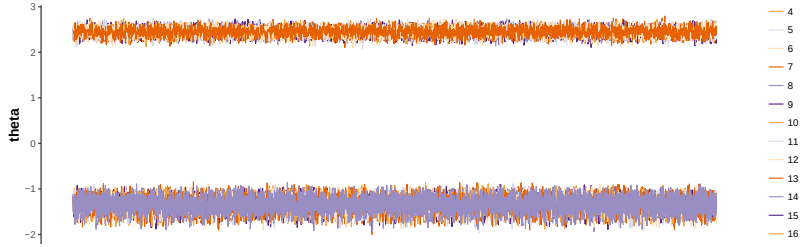
If we look at the posterior predictive distribution

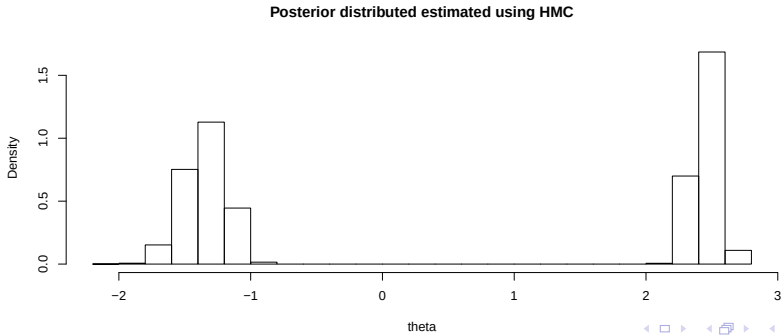
$$\pi(x|D=2) = \int \pi(x|\theta)\pi_u(\theta)\mathrm{d}\theta$$

then we get a very different answer if we propagate through the Bayesian posterior $\pi_u(\theta) = \pi(\theta|D=2)$ and the asymptotic distribution of the MLE $\pi_u(\theta) = N(\theta, \mathcal{I}^{-1})$.



Aside: HMC (in stan) 20 chains from random start points





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

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

$$e \sim U[-\epsilon, \epsilon]$$

We can generalise ABC to assume non-uniform errors

Key challenges for ABC (or perhaps for all inference)

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 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

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

MCMC-ABC

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

We are targeting the joint distribution

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

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

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

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

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)

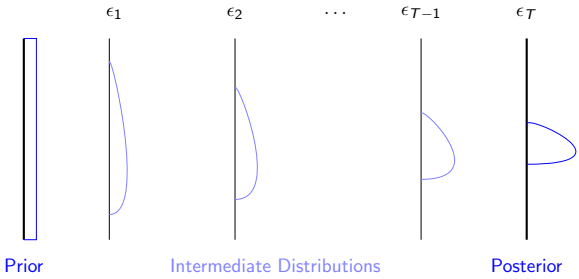
Sequential ABC algorithms

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

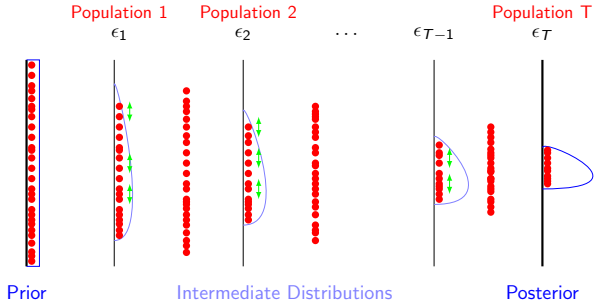
Choose a sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ and let π_t be the ABC approximation when using tolerance ϵ_t.

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

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?

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 $S(D)$ 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 $S(D)$ 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, often 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, often better than humans.

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

E.g. 1)Use random forests, (C)NNs etc to generate a summary

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, often better than humans.

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

E.g. 1)Use random forests, (C)NNs etc to generate a summary

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) 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, often better than humans.

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

E.g. 1)Use random forests, (C)NNs etc to generate a summary

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) 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 the kernel mean embedding of the distribution (MMD) in an RKHS - inference is then simply projection in the RKHS.

Machine learning invasion

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

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

E.g. 1)Use random forests, (C)NNs etc to generate a summary

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) 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 the kernel mean embedding of the distribution (MMD) in an RKHS - inference is then simply projection in the RKHS.

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

- Warning: beware of all automated summary selection approaches if misspecified

Inference for misspecified models

Inference under discrepancy

How should we do inference if the model is imperfect?
Data generating process

$$y \sim G$$

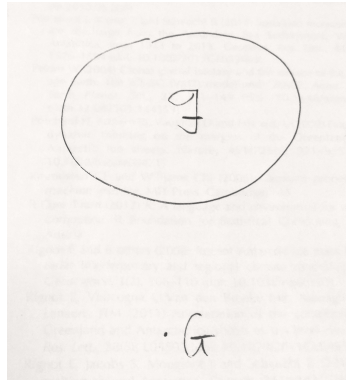
Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If $G = F_{\theta_0} \in \mathcal{F}$ then we know what to do[1].

How should we proceed if

$$G \notin \mathcal{F}$$



[1] Even if we can't agree about it!

Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} l(y|\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$, then (under some conditions)

$$\hat{\theta}_n \to \theta_0 \text{ almost surely as } \quad n \to \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, \mathcal{I}^{-1}(\theta_0))$$

Asymptotic consistency, efficiency, normality.

Maximum likelihood

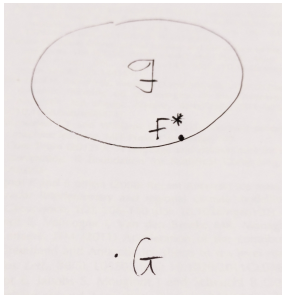
Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} l(y|\theta)$$

If $G = F_{\theta_0} \in \mathcal{F}$, then (under some conditions)

$$\hat{\theta}_n \to \theta_0 \text{ almost surely as } \quad n \to \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, \mathcal{I}^{-1}(\theta_0))$$



Asymptotic consistency, efficiency, normality.
If $G \notin \mathcal{F}$

$$\hat{\theta}_n \to \theta^* = \arg\min_{\theta} D_{KL}(G, F_\theta) \text{ almost surely}$$

$$= \arg\min_{\theta} \int \log \frac{dG}{dF_\theta} dG$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Rightarrow} N(0, V^{-1})$$

Bayes

Bayesian posterior

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

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality (under some conditions).

Bayes

Bayesian posterior

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

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality (under some conditions).

If $G \notin \mathcal{F}$, we still get asymptotic concentration (and possibly normality) but to θ^* (the pseudo-true value).

"there is no obvious meaning for Bayesian analysis in this case"

Bayes

Bayesian posterior

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

If $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality (under some conditions).

If $G \notin \mathcal{F}$, we still get asymptotic concentration (and possibly normality) but to θ^* (the pseudo-true value).

> *"there is no obvious meaning for Bayesian analysis in this case"*

Often with non-parametric models (eg GPs), we don't even get this convergence to the pseudo-true value due to lack of identifiability.

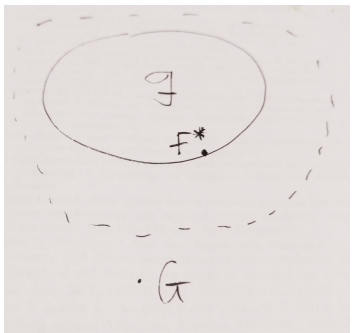
An appealing idea: model the discrepancy

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: model the discrepancy

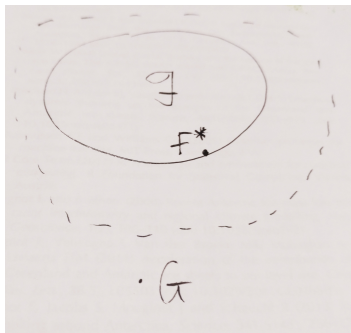
Kennedy an O'Hagan 2001

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

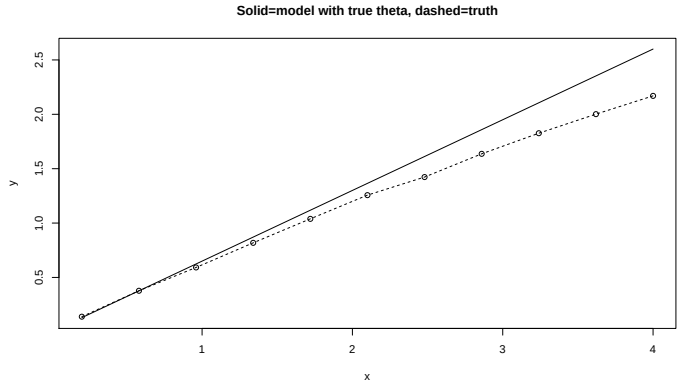
Kennedy and O'Hagan 2001, Brynjarsdottir and O'Hagan 2014

Simulator

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

Reality

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



An appealing, but flawed, idea

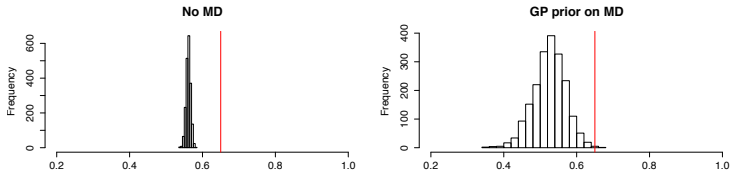
Bolting on a GP can correct your predictions, but won't necessarily fix your inference:

- No discrepancy:

$$y = f_\theta(x) + N(0, \sigma^2),$$
$$\theta \sim N(0{,}100), \sigma^2 \sim \Gamma^{-1}(0.001, 0.001)$$

- GP discrepancy:

$$y = f_\theta(x) + \delta(x) + N(0, \sigma^2),$$
$$\delta(\cdot) \sim GP(\cdot, \cdot) \text{ with objective priors}$$



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 too complex to understand

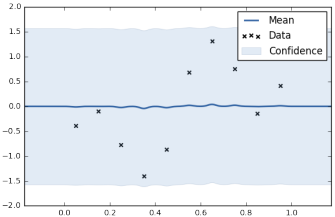
Dangers of non-parametric model extensions

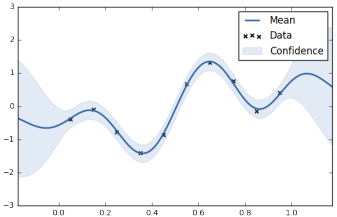
There are (at least) two problems with this approach:

- 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.

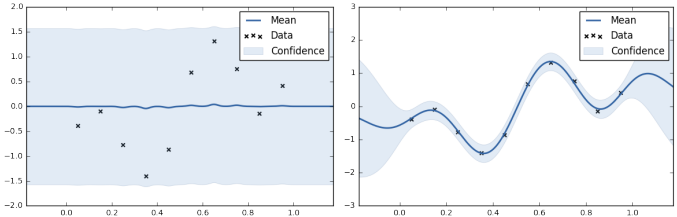
    ie We never forget the prior, but the prior is too complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.

- We can also have problems finding the true optima for the hyperparameters, even in 1d problems:





- We can also have problems finding the true optima for the hyperparameters, even in 1d problems:



- Wong et al 2017 impose identifiability (for δ and θ) by giving up and identifying

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

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 2, *pp.* 635–648

A frequentist approach to computer model calibration

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

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

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

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

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

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

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?
  - We should use methods that limit the amount of learning that is possible 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?
  - We should use methods that limit the amount of learning that is possible 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

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.

Scoring of parameter values needs careful thought

- Likelihood isn't always fit for purpose.

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.

Scoring of parameter values needs careful thought

- Likelihood isn't always fit for purpose.

Thank you for listening!