Approximate Bayesian computation: likelihood-free inference for complex models

Richard Wilkinson

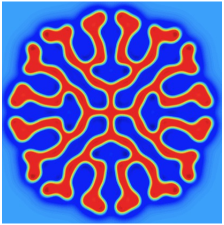
School of Maths and Statistics
University of Sheffield

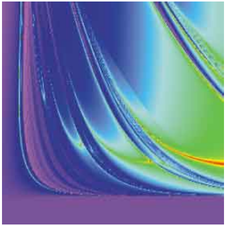
April 21, 2017

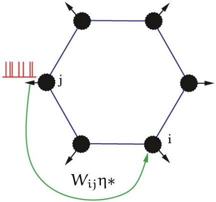
Calibration

- 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 inverse/ calibration/ parameter estimation/... problem is estimating θ that could have led to D







Statistical inference

Consider the following three parts of inference:

1. Modelling

2. Inferential framework

3. Statistical computation

Statistical inference

Consider the following three parts of inference:

1. Modelling
   - Simulator - generative model $\pi(X|\theta)$
   - Statistical model
     - prior distributions on unknown parameters, $\pi(\theta)$
     - observation error on the data, $\pi(D|X)$
     - simulator error (if its not a perfect representation of reality)
2. Inferential framework
3. Statistical computation

Statistical inference

Consider the following three parts of inference:

1. Modelling
   - Simulator - generative model $\pi(X|\theta)$
   - Statistical model
     - prior distributions on unknown parameters, $\pi(\theta)$
     - observation error on the data, $\pi(D|X)$
     - simulator error (if its not a perfect representation of reality)
2. Inferential framework
   - Classical/frequentist
   - Bayesian
   - History matching
3. Statistical computation

Statistical inference

Consider the following three parts of inference:

1. Modelling
   - Simulator - generative model $\pi(X|\theta)$
   - Statistical model
     - prior distributions on unknown parameters, $\pi(\theta)$
     - observation error on the data, $\pi(D|X)$
     - simulator error (if its not a perfect representation of reality)
2. Inferential framework
   - Classical/frequentist
   - Bayesian
   - History matching
3. Statistical computation
   - this remains hard even with increased computational resource

Inferential framework

Classical/frequentist

- Maximum likelihood

$$\hat{\theta} = \arg\max_{\theta} \pi(D|\theta)$$

  or a more ad-hoc approach

$$\hat{\theta} = \arg\min_{\theta} (\mathbb{E}(D|\theta) - D)^2$$

- Can find confidence intervals (with coverage guarantees etc)
- But for complex problems can be hard, and often we have additional information we want to include

Inferential framework

Classical/frequentist

- Maximum likelihood

$$\hat{\theta} = \arg\max_\theta \pi(D|\theta)$$

  or a more ad-hoc approach

$$\hat{\theta} = \arg\min_\theta (\mathbb{E}(D|\theta) - D)^2$$

- Can find confidence intervals (with coverage guarantees etc)
- But for complex problems can be hard, and often we have additional information we want to include

Bayesian

- Work only with probabilities (no significance, confidence, p-values)
- update beliefs in light of data and aim to find posterior distributions

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

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

- Needs a prior distribution, computation is still hard but often do able

Computational 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
- 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 popular in biological disciplines as they are

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

Rejection ABC

Uniform Rejection Algorithm

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

Rejection ABC

Uniform Rejection Algorithm

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

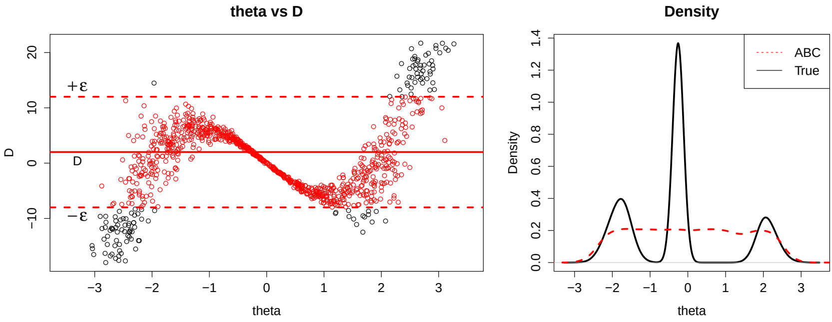
ϵ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

Rejection sampling is inefficient, but we can adapt other MC samplers such as MCMC and SMC.

Simple $\rightarrow$ Popular with non-statisticians

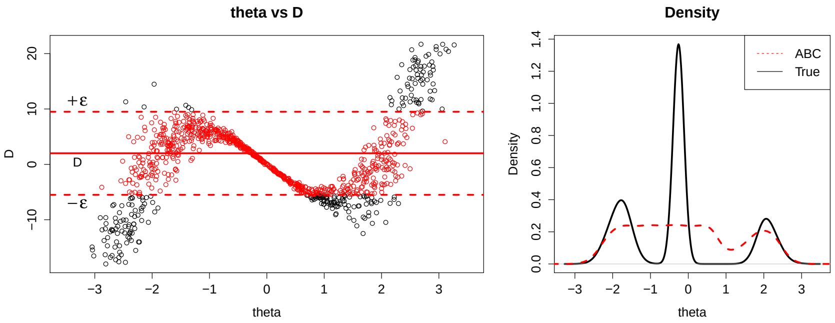
$\epsilon = 10$



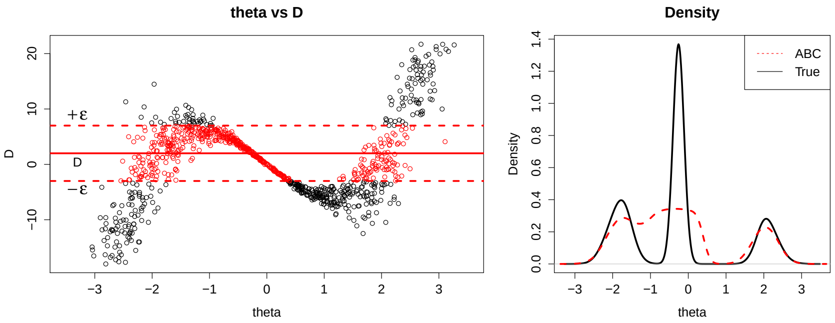
$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta+2)\theta(\theta-2), 0.1 + \theta^2)$$

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

$\epsilon = 7.5$

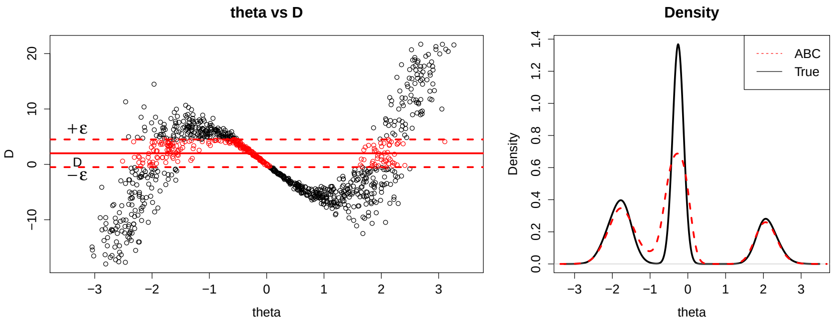

theta vs D
+ε
D
−ε
theta
D
20
10
0
−10
−3
−2
−1
0
1
2
3
Density
ABC
True
theta
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

$\epsilon = 5$



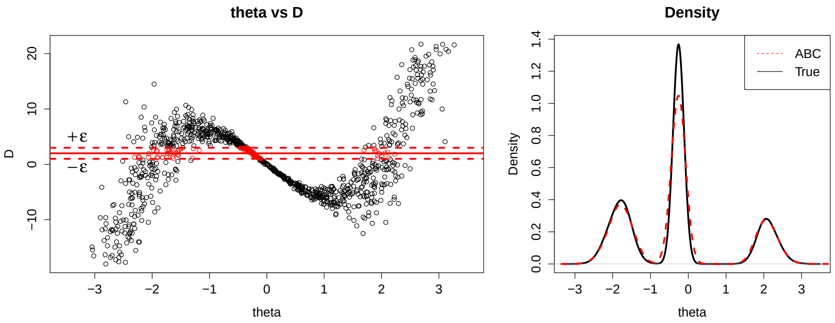
theta vs D
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3
theta
Density
ABC
True
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−3
−2
−1
0
1
2
3
theta

$\epsilon = 2.5$



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

$\epsilon = 1$



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

Rejection ABC

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

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'

- Summary statistic $S(D)$ - controls 'information loss'

Key challenges for ABC

Accuracy in ABC is determined by

- Tolerance ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ 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 ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ 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 ϵ - controls the 'ABC error'
  - how do we find efficient algorithms that allow us to use small ϵ 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

Computation

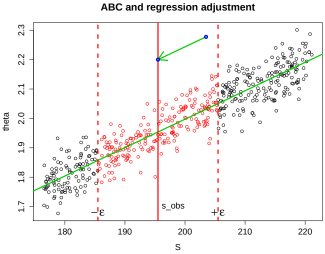
- Efficient 'exact-approximate' algorithms
  - MCMC, SMC, EM, EP, etc

Computation

- Efficient 'exact-approximate' algorithms
  - MCMC, SMC, EM, EP, etc
- Efficient 'approximate-approximate' algorithms
  - GP emulators/surrogate models
  - We can control the degree of additional approximation error here, e.g., using the surrogate to propose moves in an MCMC scheme but using the simulator to decide about acceptances.
  - Linked to Bayesian optimization

Computation

- Efficient 'exact-approximate' algorithms
  - MCMC, SMC, EM, EP, etc
- Efficient 'approximate-approximate' algorithms
  - GP emulators/surrogate models
  - We can control the degree of additional approximation error here, e.g., using the surrogate to propose moves in an MCMC scheme but using the simulator to decide about acceptances.
  - Linked to Bayesian optimization
- Post-hoc corrections



use the estimate of the posterior mean at s_{obs} and the residuals from the fitted line to form the posterior.

Error trade-off

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

1. Choice of summary:

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

Error trade-off

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

1. Choice of summary:

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

2. Use of ABC acceptance kernel:

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

Error trade-off

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

1. Choice of summary:

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

2. Use of ABC acceptance kernel:

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

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

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

1. Choice of summary:

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

2. Use of ABC acceptance kernel:

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

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 $\dim(s) = \dim(\theta)$

Choosing summary statistics

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

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?

The choice is one of the most important parts of ABC algorithms

Choosing summary statistics

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?

The choice is one of the most important parts of ABC algorithms

- Recent progress made with random forest and neural-network models to learn the relevant features
  1. Train a ML model, $m(D)$, to predict θ from D using a large number of simulator runs $\{\theta_i, D_i\}$
  2. ABC then simulates θ from the prior and D from the simulator, and accepts θ if $m(D) \approx m(D_{obs})$

Model discrepancy

- All models are wrong blah blah...

Model discrepancy

- All models are wrong blah blah...
- Doing anything about this is hard.

Model discrepancy

- All models are wrong blah blah...
- Doing anything about this is hard.
  - Appealing but often useless idea: Include a GP model of the discrepancy and infer this along with θ

Model discrepancy

- All models are wrong blah blah...
- Doing anything about this is hard.
  - Appealing but often useless idea: Include a GP model of the discrepancy and infer this along with θ
- Ignoring discrepancy can lead to over-confident and incorrect inference about θ

Model discrepancy

- All models are wrong blah blah...
- Doing anything about this is hard.
  - Appealing but often useless idea: Include a GP model of the discrepancy and infer this along with θ
- Ignoring discrepancy can lead to over-confident and incorrect inference about θ
- When using ABC, you are automatically including some characterization of model discrepancy (determined by the summaries, metric and tolerance you chose).

Model discrepancy

- All models are wrong blah blah...
- Doing anything about this is hard.
  - Appealing but often useless idea: Include a GP model of the discrepancy and infer this along with θ
- Ignoring discrepancy can lead to over-confident and incorrect inference about θ
- When using ABC, you are automatically including some characterization of model discrepancy (determined by the summaries, metric and tolerance you chose).
  - So it's better to have thought carefully about this.
  - May only be a case of thinking about an approximate magnitude of the discrepancy

History matching

History matching is a related approach usually used for complex deterministic simulators in combination with emulators.

Emphasis is less on computation and more on dealing with model discrepancy.

History matching

History matching is a related approach usually used for complex deterministic simulators in combination with emulators.

Emphasis is less on computation and more on dealing with model discrepancy.

- Find the not-implausible θ such that, e.g.,

$$\mathcal{I}(\theta) = \frac{D - \mathbb{E}(D|\theta)}{\mathbb{V}\text{ar}(D|\theta)} < 3$$

  where $\mathbb{V}\text{ar}(D|\theta)$ is the total variance taking into account measurement error, discrepancy, emulator uncertainty etc.
- Usually carried out in waves, where in each iteration more simulation is done to improve the emulator as we narrow down the plausible range of parameters.

History matching

History matching is a related approach usually used for complex deterministic simulators in combination with emulators.

Emphasis is less on computation and more on dealing with model discrepancy.

- Find the not-implausible θ such that, e.g.,

$$\mathcal{I}(\theta) = \frac{D - \mathbb{E}(D|\theta)}{\mathbb{V}\text{ar}(D|\theta)} < 3$$

  where $\mathbb{V}\text{ar}(D|\theta)$ is the total variance taking into account measurement error, discrepancy, emulator uncertainty etc.

- Usually carried out in waves, where in each iteration more simulation is done to improve the emulator as we narrow down the plausible range of parameters.

HM is a conservative approach - it only rules out parameters we are reasonably confident are implausible. It doesn't attempt to tell us the best parameter value.

Conclusions

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

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

Efficient algorithms and post-hoc regression adjustments 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.

Efficient algorithms and post-hoc regression adjustments 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.

Efficient algorithms and post-hoc regression adjustments 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!