ABC for Big Data

Paul Fearnhead

Department of Mathematics & Statistics, Lancaster University

31st January 2013

Introduction

There is increasing need to perform inference for complex stochastic models that are easy to simulate from, but for which it is difficult to calculate likelihoods for.

Can we use the fact we can simulate from these models to help us make inferences?

Yes! For example, we can estimate likelihoods by the proportion of times simulated data is similar to the observed data.

Likelihood-free Methods

Arguably these methods date back to (Diggle and Gratton 1984).

Variety of methods: Indirect Inference (Gouriéoux and Ronchetti 1993); bootstrap filter (Gordon et al. 1993); amongst many others (Wood 2010, Cox and Kartsonaki 2012).

We are considering a Bayesian approach: Approximate Bayesian Computation (ABC) (Pritchard et al. 1999, Marjoram et al. 2003).

Part I

Approximate Bayesian Computation

Approximate Bayesian Computation

It is easiest to see the idea of ABC through a simple rejection sampling algorithm.

In practice more efficient MCMC, Importance Sampling or Sequential Importance Sampling methods exist.

ABC Rejection Sampling

Input: observed data $\mathbf{x}_{\text{obs}}$, threshold $h > 0$

For $i = 1, 2, \ldots, n$:

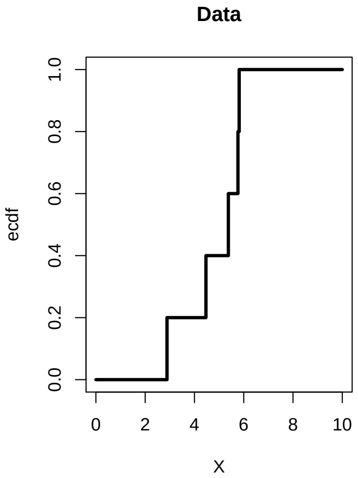
1. Sample parameter vector θ_i from prior $\pi(\theta)$;
2. Simulate data $\mathbf{x}_{\text{sim}}$ from model conditional on θ_i;
3. If $||\mathbf{x}_{\text{sim}} - \mathbf{x}_{\text{obs}}|| < h$ accept θ_i.

Output is a sample of θ values from an approximation to the posterior.

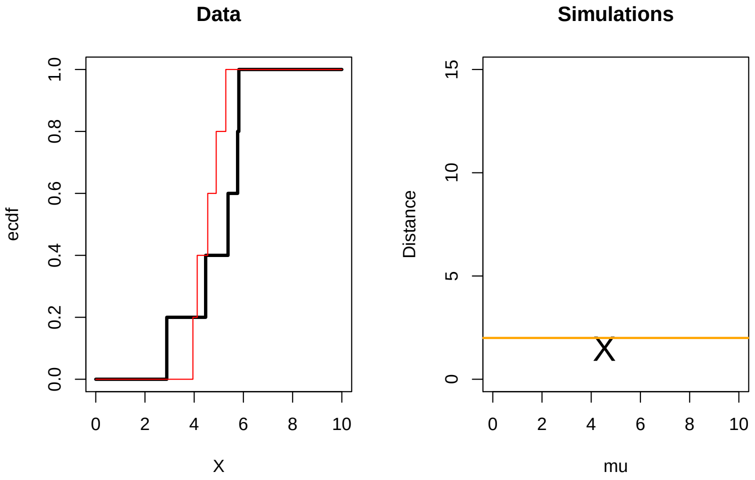
ABC: Example

- Model: 5 IID draws from $N(\mu, 1)$.
- Uniform prior on $[0, 10]$.
- Use Euclidean distance between ordered data values.

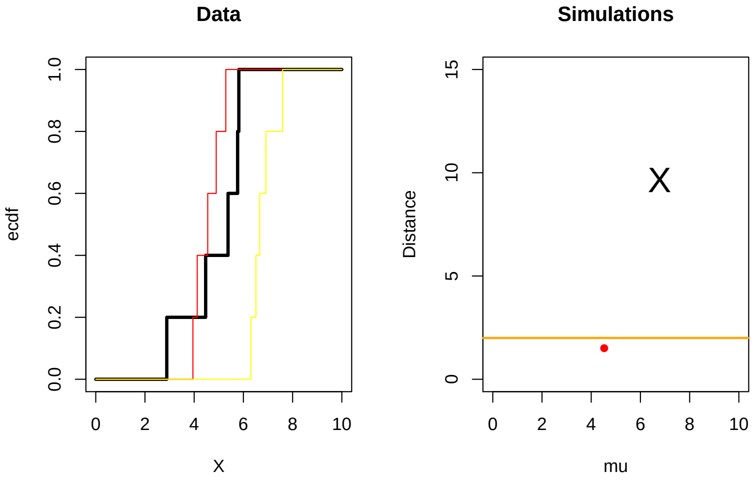
ABC: example



ABC: example

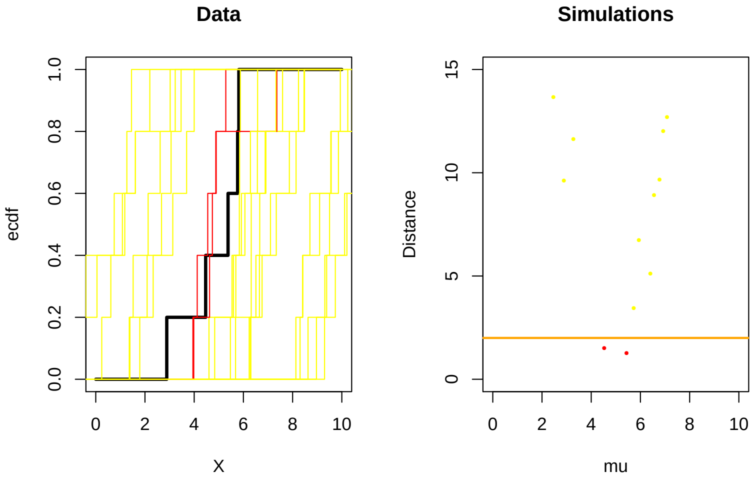

Data
ecdf
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
X
Simulations
Distance
0
5
10
15
0
2
4
6
8
10
mu
X

ABC: example

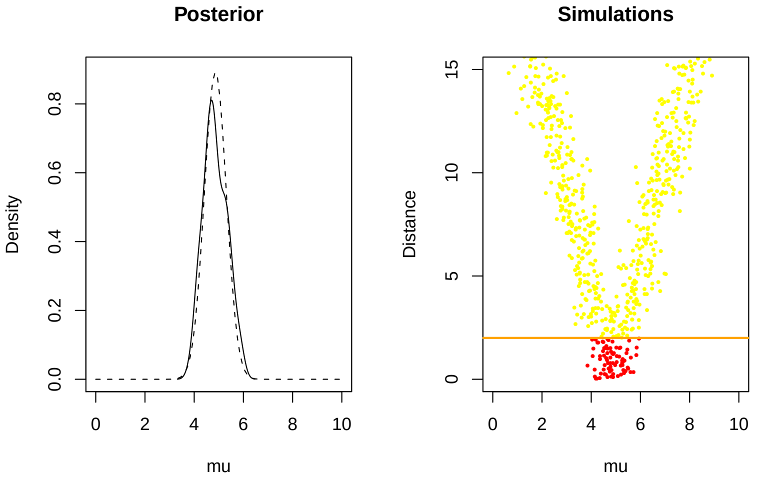


Data
ecdf
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
X
Simulations
Distance
0
5
10
15
X
0
2
4
6
8
10
mu

ABC: example


Data
ecdf
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
X
Simulations
Distance
0
5
10
15
0
2
4
6
8
10
mu

ABC: example



Posterior
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
mu
Simulations
Distance
0
5
10
15
0
2
4
6
8
10
mu

Approximations in ABC

Call the distribution the rejection sampler samples from the ABC posterior.

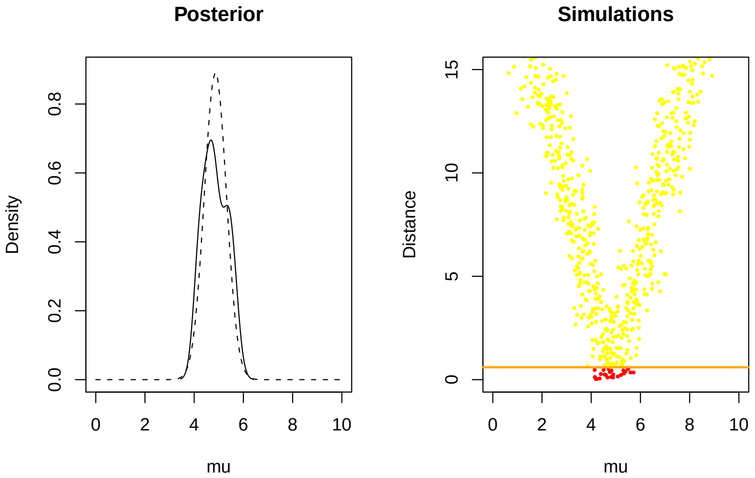
If we only accepted θ_i if

$$\mathbf{x}_{\text{sim}} = \mathbf{x}_{\text{obs}}$$

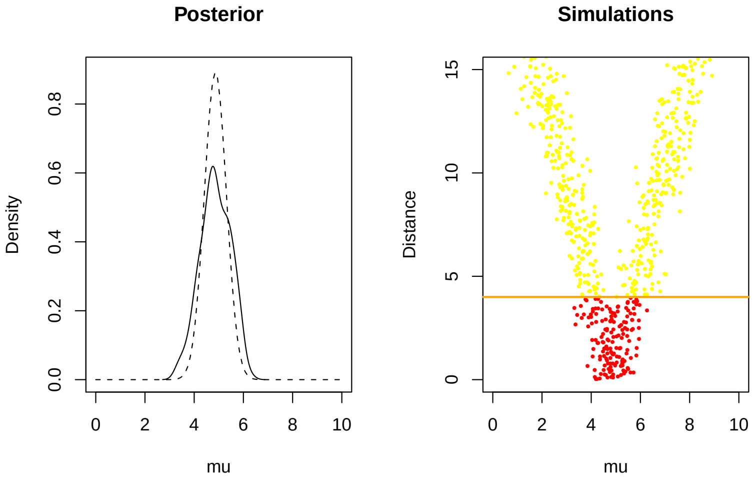
then the ABC posterior is the true posterior.

In practice accepting based on $||\mathbf{x}_{\text{sim}} - \mathbf{x}_{\text{obs}}|| < h$ introduces a trade-off between approximation error in the ABC posterior and Monte Carlo error.

ABC: example


Posterior
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
mu
Simulations
Distance
0
5
10
15
0
2
4
6
8
10
mu

ABC: example



Posterior
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
mu
Simulations
Distance
0
5
10
15
0
2
4
6
8
10
mu

Scalability of ABC

For moderate-dimensional data we are unlikely to get a close match between all the simulated and observed data sets for all data values.

To apply ABC for high-dimensional data: we need to accept if summaries of the simulated data are similar to summaries of the real-data.

How do we choose summary statistics which are informative and low-dimensional?

Different view of ABC

The standard view of ABC is that it gives an approximation to the true posterior:

$$\pi_{\text{ABC}}(\theta|\mathbf{x}_{\text{obs}}) \approx \pi(\theta|\mathbf{x}_{\text{obs}}).$$

This suggests finding approximate sufficient statistics.

This is unrealistic. Instead we should evaluate ABC in terms of its inferential properties. We want accurate estimates. (And to understand the ABC posterior?).

Part II

Accuracy: Choice of Summary Statistics

Definition of Accuracy

Consider $\phi_i = f_i(\theta)$, for $i = 1, \dots, p$. We will quantify accuracy in terms of estimating $\phi = (\phi_1, \dots, \phi_p)$.

Often we would have $\phi = \theta$.

We define this in terms of a Loss Function. We consider square error-loss:

$$L(\phi, \hat{\phi}; A) = (\phi - \hat{\phi})A(\phi - \hat{\phi})^T,$$

where A is a $p \times p$ positive-definite matrix.

Optimal Summary Statistics

Consider $h \to 0$, and define accuracy in terms of square error-loss.

An optimal choice of summary statistics, $S_i(\cdot)$ is given by for $i = 1, 2, \dots, p$:

$$S_i(\mathbf{x}) = \mathrm{E}(\phi_i | \mathbf{x}).$$

This follows from

- Given the data, $\mathbf{x}$, the optimal estimate of ϕ_i is $\mathrm{E}(\phi_i|\mathbf{x})$.
- If $S_i(\mathbf{x}) = \mathrm{E}(\phi_i|\mathbf{x})$, then $\mathrm{E}_{\text{ABC}}(\phi_i|S(\mathbf{x})) \to \mathrm{E}(\phi_i|\mathbf{x})$ as $h \to 0$.

Implementation

Our result on Summary Statistics is not of direct use: as we do not know the posterior means.

However we can use simulation to estimate the functions $E(\phi_i|\mathbf{x})$. Fearnhead and Prangle (2012) show that even using linear regression, applied to simulated samples of parameters and data sets, can produce reliable summary statistics (see also Blum et al. 2012).

Part III

ABC for Big Data

ABC for big data

How does ABC scale to big data problems? Let n be the “amount” of data, and p the number of parameters. Two natural cases:

- Large n and small p: Issue of cost of simulating data. Otherwise ABC should work well. Potential for theory about ABC estimation (if large n corresponds to increasing information).
- Large n and large p: Issue of curse-of-dimensionality if number of summaries increases with p.

Subsampling

If simulating data for large n (or calculating summaries) is slow: could resort to subsampling.

Simplest approach: use a fixed subset of the data. Unsatisfactory.

Better? Use (different) random sub-samples at each iteration. In some situations it may be equivalent to using all information in the data.

[Alternatively use emulators: see AABC idea of Buzbas and Rosenberg 2013].

Exploit Structure of Model

For some hierarchical models: apply ABC independently on sub-sets of the data (Bazin et al. 2010).

For some state-space models: use Markov structure and SMC to be able to use large numbers of summaries (Barthelmé and Chopin 2011).

However, be careful about accumulation of approximation errors (Dean et al. 2011). Also how to implement SMC for large numbers of parameters?

Composite Likelihood

We could used ideas from composite likelihood: base inference on the product of likelihoods for different subsets of the data. Use ABC independently for each subset.

Introduces further approximation (assuming subsets of the data are independent).

Particularly suitable if different subsets of data are informative about different parameters.

[See also Nott et al. 2011 and Rubio and Johansen 2013.]

Discussion

How to apply ABC to big data, particularly for models with many parameters, is an open question.

To tackle this will need using ideas from other areas of statistics: rationale of `i-like`.