Introduction to Approximate Bayesian Computation

Associate Professor Chris Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

Why Statistical Models?

Why do we need statistical models?

- Gain new insights in applied disciplines
- Test hypotheses
- Make predictions

Parameter Estimation

Statistical models typically have unknown parameters.

These models would be much more useful if we knew their values.

One approach: Estimating them from data (parameter estimation or calibration).

We also want some uncertainty quantification on these estimates.

Motivation

Practitioners wish to develop more realistic models.

However, analytical calculations with such models is not possible.

But simulation from the model might be feasible.

Can we develop a parameter estimation method that relies only on model simulation?

Cell Biology Example

Cell motility and proliferation are important parts of many biological processes (e.g. skin cancer growth, wound healing).

One way to investigate this is through a scratch assay. A ‘scratch’ is made which separates the cells. Images of the cells are taken at regular time intervals until the cells are once again in contact.

Assume images taken every 5 minutes for 12 hours (145 images)

Approximately map cells onto a rectangular lattice (binary matrix where a 1 indicates presence of a cell at a particular location).

Cell Biology Example

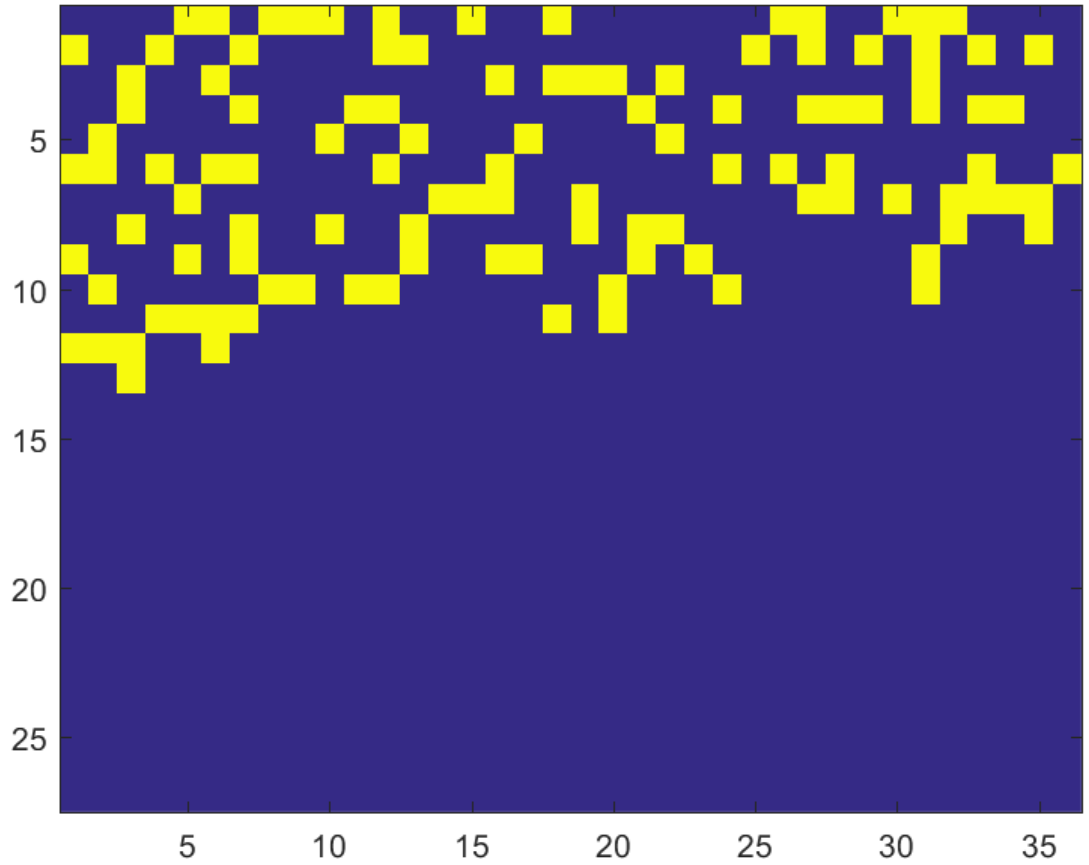
Stochastic Model (see Johnston et al 2014)

In time step τ cells given chance to move to neighbouring location with probability P_m.

During time step cells can give 'birth' with probability P_p and place new cell at neighbouring location.

Cell Biology Example

Simulated data from model with $P_m = 0.35$ and $P_p = 0.001$



Statistical Inference

Two different frameworks for parameter estimation:

- Classical
- Bayesian

Simple Example

- Consider tossing a two-sided coin 10 times with unknown probability θ of getting a head.
- Assume that in 10 tosses we get $y = 8$ heads. What is our best guess/estimate of θ?
- Intuitively, we might estimate θ as $\hat{\theta} = 8/10 = 0.8$.
- How might we get this estimate in a more statistically rigorous way.

Classical Statistics

- Denote model parameter as θ and observed data as y.
- From our model we can compute the so-called likelihood function $p(y|\theta)$ of the model. Loosely, this is the probability of getting the data for a given parameter value.
- We want the θ that maximises the likelihood function:

$$\hat{\theta} = \arg\max p(y|\theta).$$

Simple example: The number of "successes" out of a fixed number of trials has a binomial distribution. Likelihood:

$$p(y|\theta) = \binom{10}{8}\theta^8(1-\theta)^2.$$

Maximising this as a function of θ gives $\hat{\theta} = 0.8$.

Classical Statistics

The classical approach to parameter estimation is useful but...

- Uncertainty quantification of parameter estimates are based on asymptotic theory.
- Model predictions are typically based on point estimate only (no uncertainty quantification).
- Difficult to take into account prior knowledge of θ.

Bayesian Statistics

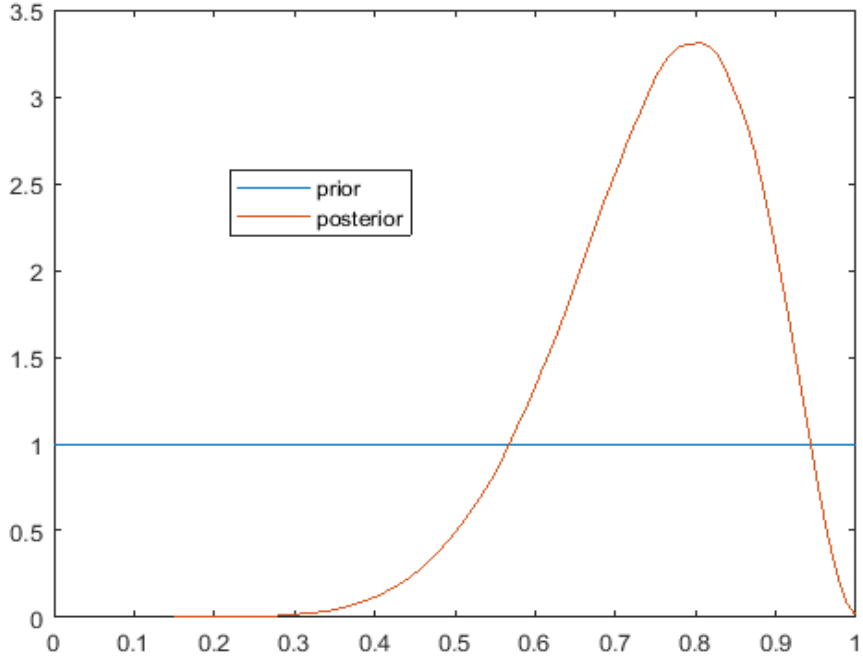
- The Bayesian approach treats θ as a random variable.
- Information about θ before data collection encapsulated in prior distribution $p(\theta)$.
- Combine with the information we obtain about θ from the data y quantified by the likelihood function. Using Bayes rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta),$$

where $p(\theta|y)$ is called the posterior distribution.

Back to Simple Example

- Consider tossing coin n times with θ as heads probability. As before, binomial likelihood with $n = 10$ and $y = 8$. Want posterior distribution of θ.
- If we have no information on θ *a priori*, we might put a uniform/flat prior on θ.
- Can analytically determine posterior.



Simple Example

Now that we have the posterior, how can we summarise it?

Point Estimate: Mean of posterior is 0.75. Mode of posterior is 0.8.

Uncertainty quantification: A 95% credible interval can be obtained via the 2.5% and 97.5% quantiles. Here this is $(0.48, 0.94)$. Compare this to classical with asymptotic assumptions (0.55, 1.05).

Advantages of Bayesian Approach

- Appropriate uncertainty quantification of parameters and predictions.
- Possible to incorporate prior knowledge. Useful for small datasets.
- Easier to interpret.
- Other advantages outside the scope of this talk.

However, the posterior is generally not available analytically nor is it possible to perfectly sample from the posterior.

Approximate Sampling from Posterior

Fortunately, there are algorithms to generate “approximate" samples from posterior.

However, standard algorithms assume that likelihood calculation is feasible.

May not be true for complex/realistic models. (such as for the motivating cell biology example)

Likelihood-Free Methods

Even in models with intractable likelihoods, it is often possible to simulate data from the model. Such as cell biology example.

For some proposed θ, likelihood-free methods simulate data x and compare with y based on some data summary $s(\cdot)$.

It is important that the chosen statistics are informative about θ. Cell biology example summary statistics:

- Sum of hamming distances between adjacent lattices (informative about P_m).
- total number of cells at end of experiment (informative about P_p).

Approximate Bayesian Computation

Approximate Bayesian computation (ABC, Sisson et al 2018) is current state-of-the-art likelihood-free Bayesian method.

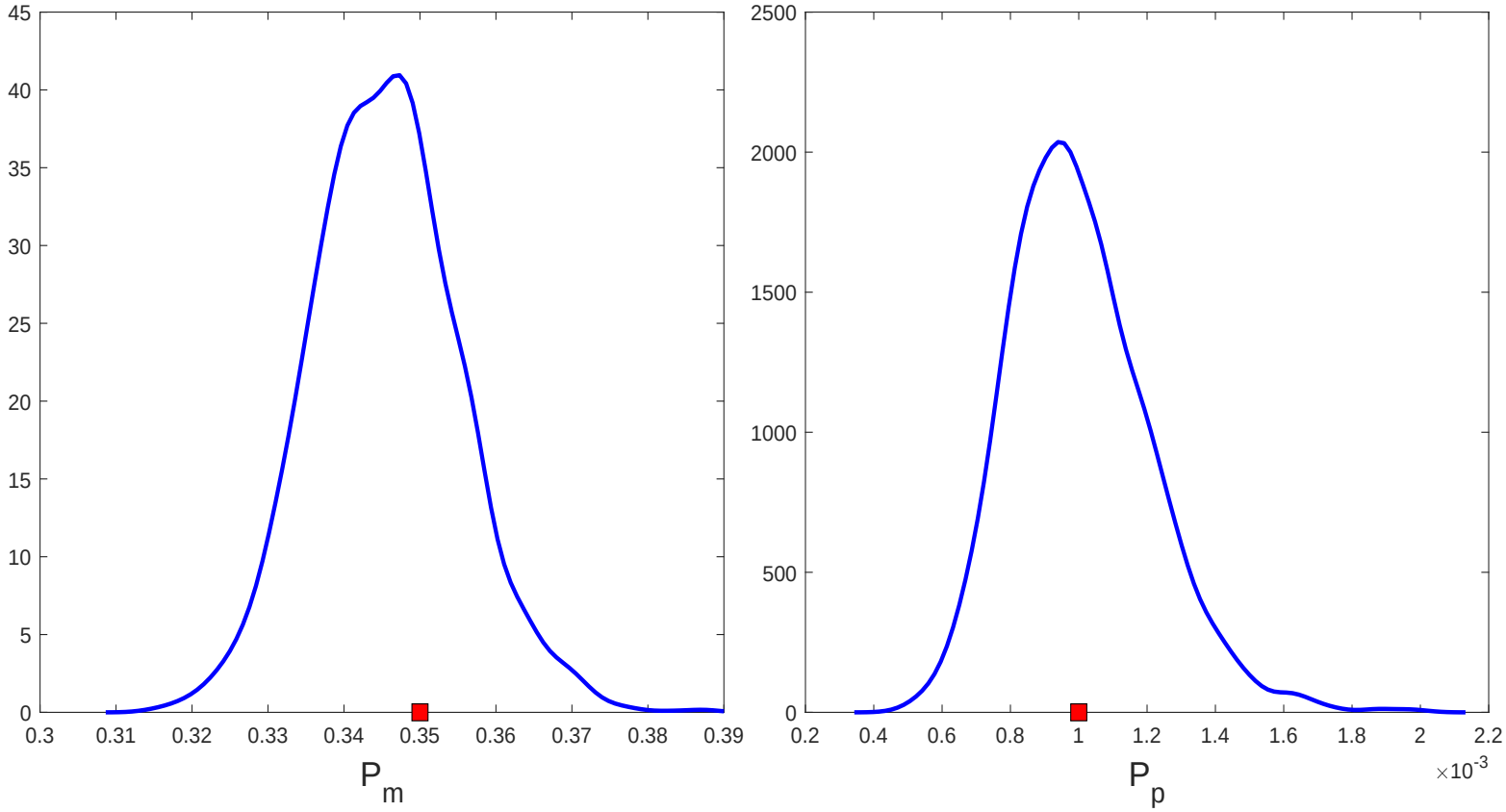
Simple concept: If a θ generates x close to y, then keep θ

'Close' typically defined if distance between $s(x)$ and $s(y)$ is below ϵ.

Caveat: ABC only provides an approximation to the true posterior.

Results for Motivating Example

Posterior Distribution



ABC Challenges

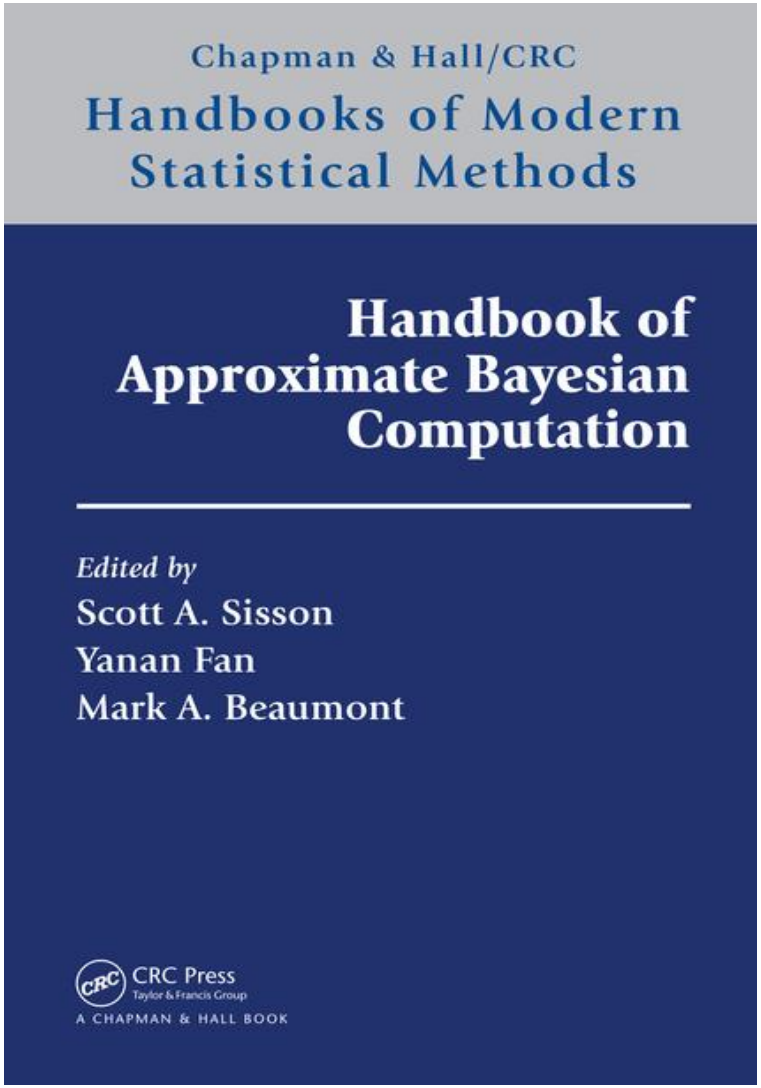
Challenges in successfully implementing ABC:

- Selecting good summary statistics.
- Selecting distance function and ϵ.
- Implementing efficient ABC algorithm for sampling posterior.

Future challenges for ABC

- High-dimensional summary statistic and/or parameter.
- Computationally expensive model simulation.
- Model selection.

First Ever Book on ABC



Thank you

Contact details:

- Email: c.drovandi@qut.edu.au
- Web: chrisdrovandi.weebly.com
- Twitter: @chris_drovandi