The *A* to *Z* of *ABC*

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

Approximate Inference Theme Day
2012

Baker 1977:

> *'Computerese is the new lingua franca of science'*

Rohrlich 1991: Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

David Cox (via Chinese whispers):

> *One of the greatest challenges facing statistics today is likelihood-free inference.*

Given a complex simulator for which we can't calculate the likelihood function - how do we do inference?

– If its cheap to simulate, then ABC (approximate Bayesian computation!) and similar ideas are the currently fashionable area of interest.

First ABC paper in 2002 (or 1999, or 1997 or ...)

By April 2010, over 250 papers developing ABC methods.

Popularity in genetics and other biological disciplines seems set to continue growing.

Talk Plan

I'll say a little on each of what I feel are the main research directions in ABC:

1. Algorithms etc
2. Regression adjustment ideas
3. Errors in ABC and the link to the computer experiment literature.

Algorithms and basics

‘Likelihood-Free’ Inference

Rejection Algorithm

- Draw θ from prior $\pi(\cdot)$
- Accept θ with probability $\pi(\mathcal{D} \mid \theta)$

Accepted θ are independent draws from the posterior distribution, $\pi(\theta \mid \mathcal{D})$.

‘Likelihood-Free’ Inference

Rejection Algorithm

- Draw θ from prior $\pi(\cdot)$
- Accept θ with probability $\pi(\mathcal{D} \mid \theta)$

Accepted θ are independent draws from the posterior distribution, $\pi(\theta \mid \mathcal{D})$.
If the likelihood, $\pi(\mathcal{D}|\theta)$, is unknown:

‘Mechanical’ Rejection Algorithm

- Draw θ from $\pi(\cdot)$
- Simulate $X \sim f(\theta)$ from the computer model
- Accept θ if $\mathcal{D} = X$, i.e., if computer output equals observation

The acceptance rate is $\mathbb{P}(\mathcal{D})$: the number of runs to get n observations is negative binomial, with mean $\frac{n}{\mathbb{P}(\mathcal{D})}$: $\Rightarrow$ Bayes Factors!

Rejection ABC

If $\mathbb{P}(\mathcal{D})$ is small, we will rarely accept any θ. Instead, there is an approximate version:

Approximate Rejection Algorithm

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

Rejection ABC

If $\mathbb{P}(\mathcal{D})$ is small, we will rarely accept any θ. Instead, there is an approximate version:

Approximate Rejection Algorithm

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

This generates observations from $\pi(\theta \mid \rho(\mathcal{D}, X) < \epsilon)$:

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

ϵ reflects the tension between computability and accuracy.

For reasons that will become clear later, we call this *uniform-ABC*.

Rejection ABC

If the data are too high dimensional we never observe simulations that are 'close' to the field data.
Reduce the dimension using summary statistics, $S(\mathcal{D})$.

Approximate Rejection Algorithm With Summaries

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

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.

Reduce the dimension using summary statistics, $S(\mathcal{D})$.

Approximate Rejection Algorithm With Summaries

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

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

Simple → Popular with non-statisticians

Embarassingly parallelizable → cluster and GPU computing

ABCifying Monte Carlo methods

Rejection ABC is the basic ABC algorithm.

A large number of papers have been published turning other MC algorithms into ABC type algorithms for when we don't know the likelihood: IS, MCMC, SMC, EM, EnKF etc

Most popular are those based on sequential methods (Sisson *et al.* 2007, Toni *et al.* 2008).

Start with the sequential importance sampler of Del Moral et al. (2006) which aims to sample a selection of N particles successively from a sequence of distributions

$$\pi_1(\theta),\ \ldots,\ \pi_T(\theta) = \text{target}$$

At stage t we aim to find a weighted cloud of particles

$$\{(\theta_i, w_i)\}_{i=1}^N$$

which approximates π_t, i.e.,

$$\pi_t(\theta) \approx \sum_{i=1}^N w_i \delta_{\theta_i}(\theta)$$

where $\delta(\cdot)$ is the Dirac delta function.

At stage t we aim to find a weighted cloud of particles

$$\{(\theta_i, w_i)\}_{i=1}^N$$

which approximates π_t, i.e.,

$$\pi_t(\theta) \approx \sum_{i=1}^N w_i \delta_{\theta_i}(\theta)$$

where $\delta(\cdot)$ is the Dirac delta function.

π_1 will typically be an easy distribution to sample from (the prior say) and we progress down (think temperature) until we reach the target distribution $\pi_T(\theta) = \pi(\theta|\mathcal{D})$.

Note that in the original conception of the particle filter the sequence of distributions is the sequence of filtering distributions

$$\begin{aligned}\pi_1 &= \pi(x_1|y_1)\\ &\vdots\\ \pi_T &= \pi(x_1,\ldots,x_T|y_1,\ldots,y_T)\end{aligned}$$

Note that in the original conception of the particle filter the sequence of distributions is the sequence of filtering distributions

$$\begin{aligned} \pi_1 &= \pi(x_1|y_1) \\ &\vdots \\ \pi_T &= \pi(x_1, \ldots, x_T|y_1, \ldots, y_T) \end{aligned}$$

This is **not** what is being done here. The sequence π_i is an arbitrary sequence of distributions choosen to provide an easy path to the target $\pi(\theta|\mathcal{D})$.

Note that in the original conception of the particle filter the sequence of distributions is the sequence of filtering distributions

$$\begin{aligned} \pi_1 &= \pi(x_1|y_1) \\ &\vdots \\ \pi_T &= \pi(x_1, \ldots, x_T|y_1, \ldots, y_T) \end{aligned}$$

This is **not** what is being done here. The sequence π_i is an arbitrary sequence of distributions choosen to provide an easy path to the target $\pi(\theta|\mathcal{D})$.

To adapt this to an ABC setting, we decide upon a sequence of tolerances

$$\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$$

and let π_t be the ABC distribution found by the ABC algorithm when we use tolerance ϵ_t.

- Think of the sequence π_i as a sequence of heated posteriors.

Toni *et al.* (2008)

Assume we have a cloud of weighted particles $\{(\theta_i, w_i)\}_{i=1}^N$ that were accepted at step $t-1$.

1. Sample θ from the previous population according to the weights.
2. Perturb the particles according to perturbation kernel q_t. I.e.,

   $$\tilde{\theta} \sim q_t(\theta, \cdot)$$

3. Reject particle immediately if $\tilde{\theta}$ has zero prior density, i.e., if

   $$\pi(\tilde{\theta}) = 0$$

4. Otherwise simulate $X \sim f(\tilde{\theta})$ from the simulator. If $\rho(S(X), S(\mathcal{D})) \leq \epsilon_t$ accept the particle, otherwise reject.
5. Give the accepted particle weight

   $$w_i = \frac{\pi(\tilde{\theta})}{\sum_{\theta_i} q_t(\theta_i, \tilde{\theta})}$$

6. Repeat steps 1-5 until we have N accepted particles at step t.

The focus is on finding more efficient algorithms, that allow us to

- do the same thing quicker, or to reduce the tolerance ϵ
- or without thinking so hard, (automating the choice of tolerance, the choice of summary statistic, the metric)

This is probably still the main focus of most ABC research.

There is also a growing literature on ABC approaches to model selection, via the approximation of Bayes Factors. There is some debate as to whether this is a good idea or not.

Regression Adjustment

Regression Adjustment

Another strand of research is based on the regression adjustment algorithms of Mark Beaumont and David Balding (Beaumont *et al.* 2002 and subsequent).

Most people tend to either go for fancy algorithms and not bother with regression adjustments

or, use rejection ABC and rely on regression adjustments after all the computation is done.

Basic idea 1

Rej-ABC produces accepted pairs $\{(\theta_i, s_i)\}_i$.

- Weight each pair by $K_\epsilon(\rho(s_{obs}, s_i))$ for some kernel $K(\cdot)$ (e.g. Epanechnikov)

If we want to estimate $\mathbb{E}(g(\theta)|s_{obs})$,

$$\mathbb{E}(g(\theta)|s_{obs}) = \int g(\theta)\pi(\theta|s_{obs})\mathrm{d}\theta \qquad (1)$$

$$= \int g(\theta)\frac{\pi(\theta, s_{obs})}{\pi(s_{obs})} \qquad (2)$$

Approximate $\pi(\theta, s_{obs})$ and $\pi(s_{obs})$ using kernel density estimates

$$\hat{\pi}(\theta, s_{obs}) = \frac{1}{n}\sum_i K_\epsilon(\rho(s_{obs}, s_i))\theta_i \qquad \hat{\pi}(s_{obs}) = \frac{1}{n}\sum_i K_\epsilon(\rho(s_{obs}, s_i))$$

and substitute to get the Nadaraya-Watson estimator:

$$\mathbb{E}(g(\theta)|s_{obs}) \approx \frac{\sum_i K_\epsilon(\rho(s_{obs}, s_i))\theta_i}{\sum_i K_\epsilon(\rho(s_{obs}, s_i))}$$

Basic idea 2

Assume a linear model for the posterior

$$\theta_i = a + (s_i - s_{obs})^T b + e_i$$
$$\mathbb{E}(\theta|s_i) = a + (s_i - s_{obs})^T b$$

$\hat{a}$ is an estimate of $\mathbb{E}(\theta|s_{obs})$, and with the empirical residuals $\hat{e} = \theta_i - \hat{a} - (s_i - s_{obs})^T \hat{b}$ form an approximation to the posterior

$$\pi(\theta|s_{obs}) \approx \hat{a} + \sum_i \delta_{\hat{\theta}_i}(\theta) \quad \text{where } \hat{\theta}_i = \hat{a} + \hat{e} \text{ are the fitted values}$$

Linear assumptions will usually be false globally, but often hold in the vicinity of s_{obs}.

- use local linear regression and weight (θ_i, s_i) by $K_\epsilon(\rho(s_i, s_{obs}))$ and minimize

$$\sum (\theta_i - a - (s_i - s_{obs})^T b)^2 K_\epsilon(\rho(s_i, s_{obs}))$$

There are many extensions - including using GP regression.

ABC and the computer experiment literature

Calibration framework

Kennedy and O'Hagan 2001

Writing $\pi(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$ can be misleading, as $\pi(\mathcal{D}|\theta)$ is not just the simulator likelihood function.

The usual way of thinking of the calibration problem is

- Relate the best-simulator run $(X = f(\hat{\theta}, t))$ to reality $\zeta(t)$
- Relate reality to the observations.

$$\begin{tikzcd}
\hat{\theta} \arrow[r] & f(\hat{\theta}, t) \arrow[r] & \zeta(t) \arrow[r] & D(t) \\
 & \text{simulator error} \arrow[ur] & \text{measurement error} \arrow[ur] &
\end{tikzcd}$$

Calibration framework

Mathematically, we can write the likelihood as

$$\pi(D|\theta) = \int \pi(D|x)\pi(x|\theta)\mathrm{dx}$$

where

- $\pi(D|x)$ is a pdf relating the simulator output to reality - the *acceptance kernel*.
- $\pi(x|\theta)$ is the likelihood function of the simulator (ie not relating to reality)

Calibration framework

Mathematically, we can write the likelihood as

$$\pi(D|\theta) = \int \pi(D|x)\pi(x|\theta)\mathrm{dx}$$

where

- $\pi(D|x)$ is a pdf relating the simulator output to reality - the *acceptance kernel*.
- $\pi(x|\theta)$ is the likelihood function of the simulator (ie not relating to reality)

Working in joint (θ, x) space, we then find

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

NB: we can allow $\pi(D|X)$ to depend on (part of) θ.

Acceptance Kernel - $\pi(D|x)$

How do we relate the simulator to reality?

1. Measurement error - $D = \zeta + e$ - let $\pi(D|X) = \pi(D - X)$ be the distribution of measurement error e.

Acceptance Kernel - $\pi(D|x)$

How do we relate the simulator to reality?

1. Measurement error - $D = \zeta + e$ - let $\pi(D|X) = \pi(D - X)$ be the distribution of measurement error e.
2. Model error - $\zeta = f(\theta) + \epsilon$ - let $\pi(D|X) = \pi(D - X)$ be the distribution of the model error ϵ.

   Kennedy and O'Hagan & Goldstein and Rougier used model and measurement error, which makes $\pi(D|x)$ a convolution of the two distributions (although they simplified this by making Gaussian assumptions).

Acceptance Kernel - $\pi(D|x)$

How do we relate the simulator to reality?

1. Measurement error - $D = \zeta + e$ - let $\pi(D|X) = \pi(D - X)$ be the distribution of measurement error e.
2. Model error - $\zeta = f(\theta) + \epsilon$ - let $\pi(D|X) = \pi(D - X)$ be the distribution of the model error ϵ.

   Kennedy and O'Hagan & Goldstein and Rougier used model and measurement error, which makes $\pi(D|x)$ a convolution of the two distributions (although they simplified this by making Gaussian assumptions).
3. Sampling of a hidden space - often the data D are simple noisy observations of some latent feature (call it X), which itself is generated by a stochastic process. By removing the stochastic sampling from the simulator we can let $\pi(D|x)$ do the sampling for us (Rao-Blackwellisation).

How does ABC relate to Kennedy's calibration ideas?

Wilkinson 2008

The distribution obtained from ABC is usually denoted

$$\pi(\theta|\rho(D,X)\leq\delta)$$

This notation is unhelpful.

How does ABC relate to Kennedy's calibration ideas?

Wilkinson 2008

The distribution obtained from ABC is usually denoted

$$\pi(\theta|\rho(D, X) \leq \delta)$$

This notation is unhelpful.

Instead, write down the ABC distribution (again in joint space):

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

How does ABC relate to Kennedy's calibration ideas?

Wilkinson 2008

The distribution obtained from ABC is usually denoted

$$\pi(\theta|\rho(D,X) \leq \delta)$$

This notation is unhelpful.

Instead, write down the ABC distribution (again in joint space):

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

We can now instantly see the relatonship between ABC and the calibration framework outlined earlier:

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

How does ABC relate to Kennedy's calibration ideas?

Wilkinson 2008

The distribution obtained from ABC is usually denoted

$$\pi(\theta|\rho(D,X) \leq \delta)$$

This notation is unhelpful.

Instead, write down the ABC distribution (again in joint space):

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

We can now instantly see the relatonship between ABC and the calibration framework outlined earlier:

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

If we replace the indicator function $\mathbb{I}_{\rho(D,x)\leq\epsilon}$ in ABC by a general acceptance kernel $\pi(D|X)$, we gain control of the approximaton.

Generalized ABC (GABC)

The rejection algorithm then becomes

Generalised Approximate Rejection Algorithm

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$ (ie $(\theta, X) \sim g(\cdot)$)
2. Accept (θ, X) if

$$U \sim U[0,1] \leq \frac{\pi(D|X)}{\max_x \pi(D|x)}$$

Generalized ABC (GABC)

The rejection algorithm then becomes

Generalised Approximate Rejection Algorithm

1. $\theta \sim \pi(\theta)$ and $X \sim \pi(x|\theta)$ (ie $(\theta, X) \sim g(\cdot)$)
2. Accept (θ, X) if

$$U \sim U[0,1] \leq \frac{\pi(D|X)}{\max_x \pi(D|x)}$$

In uniform ABC we take

$$\pi(D|X) = \begin{cases} 1 & \text{if } \rho(D,X) \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

this reduces the algorithm to

2' Accept θ ifF $\rho(D,X) \leq \epsilon$

ie, we recover the *uniform* ABC algorithm.

Uniform ABC algorithm

This allows us to interpret uniform ABC. Suppose $X, D \in \mathcal{R}$

Proposition

Accepted θ from the uniform ABC algorithm (with $\rho(D, X) = |D - X|$) are samples from the posterior distribution of θ given D where we assume $D = f(\theta) + e$ and that

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

In general, uniform ABC assumes that

$$D|x \sim U\{d : \rho(d, x) \leq \epsilon\}$$

We can think of this as assuming a uniform error term when we relate the simulator to the observations.

Uniform ABC algorithm

This allows us to interpret uniform ABC. Suppose $X, D \in \mathcal{R}$

Proposition

Accepted θ from the uniform ABC algorithm (with $\rho(D, X) = |D - X|$) are samples from the posterior distribution of θ given D where we assume $D = f(\theta) + e$ and that

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

In general, uniform ABC assumes that

$$D|x \sim U\{d : \rho(d, x) \leq \epsilon\}$$

We can think of this as assuming a uniform error term when we relate the simulator to the observations.

ABC gives 'exact' inference under a different model!

GABC Extensions

This framework can be (and has been) extended to all the other forms of ABC, eg MCMC, SMC etc (forthcoming).

GABC allows us to

- generalise ABC algorithms to move beyond the use of uniform error structures and use the added variation to include information about the error on the data and in the model.
- Improve efficiency as smoother acceptance kernels can lead to better mixing than can be achieved using the step function implied by uniform-ABC.

Oli Ratmann (Duke, but soon to be Imperial) has a sequence of papers on using ABC for model criticism.

The dangers of ABC - H.L. Mencken

> *For every complex problem, there is an answer that is short, simple and wrong*

Why use ABC? J. Galsworthy

> *Idealism increases in direct proportion to ones distance from the problem*