Exploring the error in ABC algorithms

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

Warwick - November 2010

Talk Plan

1. Brief intro to computer experiments
2. Current ABC algorithms
3. Generalised ABC algorithms
4. Examples

Computer experiments

Baker 1977 (Science):

> *'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)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

Computer experiments

Baker 1977 (Science):

> *'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)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

- how do we relate simulators to reality? (model error)
- how do we estimate tunable parameters? (calibration)
- how do we deal with computational constraints? (stat. comp.)
- how do we make uncertainty statements about the world that combine models, data and their corresponding errors? (UQ)

There is an inherent a lack of quantitative information on the uncertainty surrounding a simulation - unlike in physical experiments.

Calibration

Focus on simulator calibration:

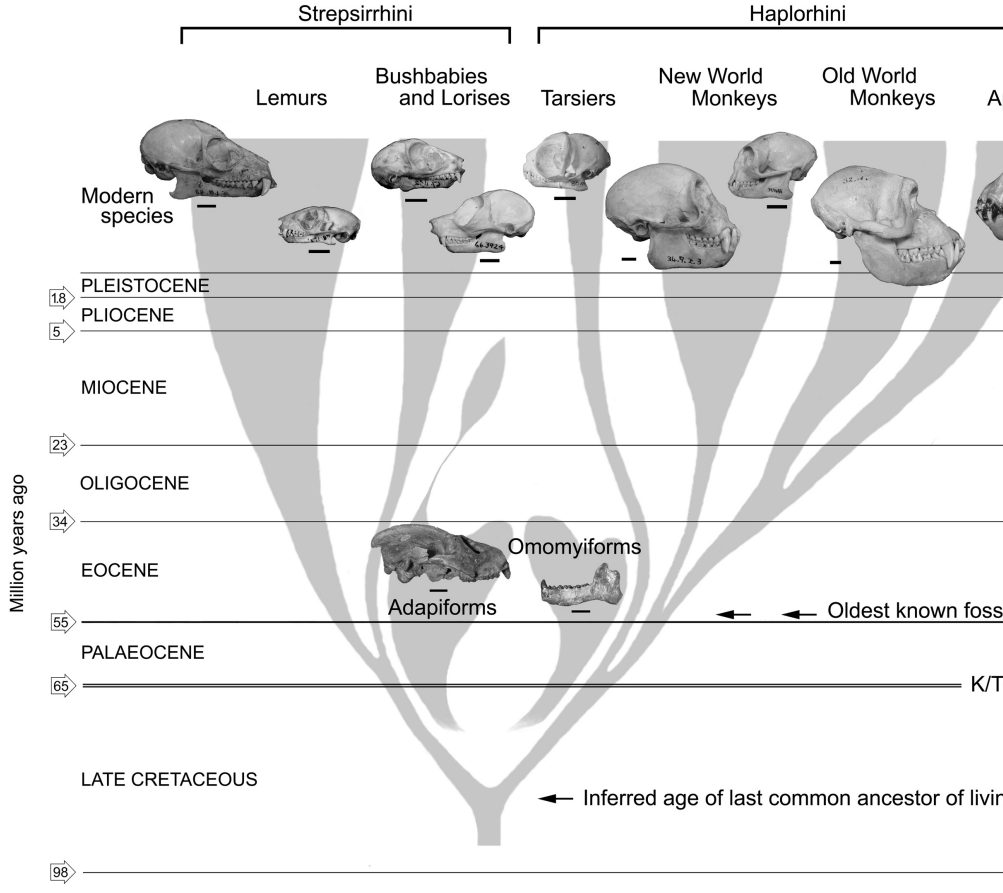
- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- We are interested in the inverse-problem, i.e., observe data D, want to estimate parameter values θ which explain this data.

For Bayesians, this is a question of fir the posterior distribution

$$\pi(\theta|\mathcal{D}) \propto; \pi(\theta)\pi(\mathcal{D}|\theta)$$

posterior $\propto$

prior $\times$ likelihd



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
   - Statistical model - priors on unknown parameters, observation error on the data, 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
   - Statistical model - priors on unknown parameters, observation error on the data, simulator error (if its not a perfect representation of reality)
2. Inferential framework - Bayesian: update beliefs in light of data and aim to find posterior distributions

$$\pi(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

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

   Note: the posterior depends on all of the modelling choices

3. Statistical computation

Statistical inference

Consider the following three parts of inference:

1. Modelling
   - Simulator - generative model
   - Statistical model - priors on unknown parameters, observation error on the data, simulator error (if its not a perfect representation of reality)
2. Inferential framework - Bayesian: update beliefs in light of data and aim to find posterior distributions

$$\pi(\theta|\mathcal{D}) \propto \pi(\theta)\pi(\mathcal{D}|\theta)$$

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

   Note: the posterior depends on all of the modelling choices

3. Statistical computation - this remains hard even with increased computational resource

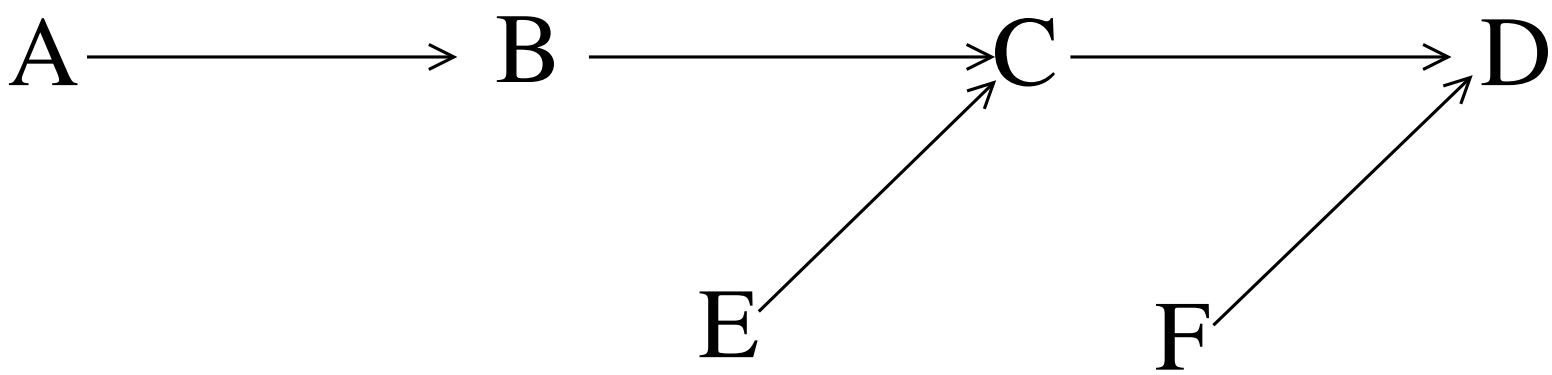
The existence of model or measurement error can make the specification of both the prior and likelihood challenging.

Calibration framework

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.



See, for example, Kennedy and O'Hagan (2001, Ser. B) & Goldstein and Rougier (2009, JSPI).

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)

This gives the desired posterior to be

$$\pi(\theta|D) = \frac{1}{Z}\int \pi(D|x)\pi(x|\theta)\mathrm{dx}.\ \pi(\theta)$$

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

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)

This gives the desired posterior to be

$$\pi(\theta|D) = \frac{1}{Z}\int \pi(D|x)\pi(x|\theta)\mathrm{dx}.\ \pi(\theta)$$

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

To simplify matters, we can work in joint (θ, x) space

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

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).

Approximate Bayesian Computation (ABC)

Approximate Bayesian computation (ABC) algorithms are a collection of Monte Carlo algorithms used for calibrating simulators

- they do not require explicit knowledge of the likelihood function $\pi(x|\theta)$
- instead, inference is done using simulation from the model (consequently they are sometimes called 'likelihood-free').

ABC methods have become popular in the biological sciences.

Although their current statistical incarnation originates from a 1999 paper (Pritchard *et al.*), heuristic versions of the algorithm exist in most modelling communities.

Uniform Approximate Bayesian Computation Algorithms

Uniform ABC

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

For reasons that will become clear later, we shall call this *Uniform ABC*.

Uniform Approximate Bayesian Computation Algorithms

Uniform ABC

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

For reasons that will become clear later, we shall call this *Uniform ABC*.

- As $\delta \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\delta = 0$, we generate observations from $\pi(\theta \mid \mathcal{D},\ \text{PMH})$ (where PMH signifies that we have made a perfect model hypothesis - no model or measurement error - unless it is simulated).

δ reflects the tension between computability and accuracy.

How does ABC relate to calibration?

The distribution obtained from ABC is usually denoted

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

This notation is unhelpful.

The hope is that $\pi(\theta|\rho(D,X) \leq \delta) \approx \pi(\theta|D,\ \text{PMH})$ for δ small.

How does ABC relate to calibration?

The distribution obtained from ABC is usually denoted

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

This notation is unhelpful.

The hope is that $\pi(\theta|\rho(D,X) \leq \delta) \approx \pi(\theta|D,\ \text{PMH})$ for δ small.

Instead, lets aim to understand the approximation, control it, and make the most of it.

To do this we can think about how the algorithm above relates to the calibration framework outlined earlier:

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

Generalized ABC (GABC)

Consider simulating from the target distribution

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

Lets sample from this using the rejection algorithm with instrumental distribution

$$g(\theta, x) = \pi(x|\theta)\pi(\theta)$$

and note that $supp(\pi_{ABC}) \subseteq supp(g)$ and that there exists a constant $M = \frac{\max_x \pi(D|X)}{Z}$ such that

$$\pi_{ABC}(\theta, x) \leq Mg(\theta, x) \qquad \forall\, (\theta, x)$$

Generalized ABC (GABC)

The rejection algorithm then becomes

Approximate Rejection Algorithm With Summaries

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

$$U \sim U[0,1] \leq \frac{\pi_{ABC}(\theta, x)}{Mg(\theta, x)} = \frac{\pi(D|X)}{MZ_{ABC}} = \frac{\pi(D|X)}{\max_x \pi(D|x)}$$

Generalized ABC (GABC)

The rejection algorithm then becomes

Approximate Rejection Algorithm With Summaries

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

$$U \sim U[0,1] \leq \frac{\pi_{ABC}(\theta, x)}{Mg(\theta, x)} = \frac{\pi(D|X)}{MZ_{ABC}} = \frac{\pi(D|X)}{\max_x \pi(D|x)}$$

In uniform ABC we take

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

this reduces the algorithm to

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

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 $D = f(\theta) + \epsilon$ and that

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

In general, uniform ABC assumes that

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

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 $D = f(\theta) + \epsilon$ and that

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

In general, uniform ABC assumes that

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

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!

Advantages of GABC

GABC

- allows us to make the inference we want to make
  - - makes explicit the assumptions about the relationship between simulator and observations.
- allows for the possibility of more efficient ABC algorithms
  - - the 0-1 uniform cut-off is less flexible and forgiving than using generalised kernels for $\pi(D|X)$
- allows for new ABC algorithms, as (non-trivial) importance sampling algorithms are now possible.
- allows us to interpret the results of ABC

Importance sampling GABC

In uniform ABC, importance sampling simply reduces to the rejection algorithm with a fixed budget for the number of simulator runs.

But for GABC it opens new algorithms:

GABC - Importance sampling

1. $\theta_i \sim \pi(\theta)$ and $X_i \sim \pi(x|\theta_i)$.
2. Give (θ_i, x_i) weight $w_i = \pi(D|x_i)$.

Importance sampling GABC

In uniform ABC, importance sampling simply reduces to the rejection algorithm with a fixed budget for the number of simulator runs.

But for GABC it opens new algorithms:

GABC - Importance sampling

1. $\theta_i \sim \pi(\theta)$ and $X_i \sim \pi(x|\theta_i)$.
2. Give (θ_i, x_i) weight $w_i = \pi(D|x_i)$.

Which is more efficient - IS-GABC or Rej-GABC?

Proposition 2

IS-GABC has a larger effective sample size than Rej-GABC, or equivalently

$$\mathbb{V}\text{ar}_{\text{Rej}}(w) \geq \mathbb{V}\text{ar}_{\text{IS}}(w)$$

This can be seen as a Rao-Blackwell type result.

Rejection Control (RC)

A difficulty with IS algorithms is that they can require the storage of a large number of particles with small weights.

A solution is to thin particles with small weights using rejection control:

Rejection Control in IS-GABC

1. $\theta_i \sim \pi(\theta)$ and $X_i \sim \pi(X|\theta_i)$
2. Accept (θ_i, X_i) with probability

$$r(X_i) = \min\left(1, \frac{\pi(D|X_i)}{C}\right)$$

   for any threshold constant $C \geq 0$.
3. Give accepted particles weights

$$w_i = \max(\pi(D|X_i), C)$$

IS is more efficient than RC, unless we have memory constraints (relative to processor time). Note that for uniform-ABC, RC is pointless.

MCMC-GABC

We can also write down a Metropolis-Hastings kernel for exploring parameter space, generalising the uniform MCMC-ABC algorithm of Marjoram *et al.*

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 (q arbitrary).

This gives the following MH kernel

MH-GABC

1. Propose a move from $z_t = (\theta, X)$ to (θ', X') using proposal Q above.
2. Accept move with probability

$$r((\theta, X), (\theta', X')) = \min\left(1, \frac{\pi(D|X')q(\theta', \theta)\pi(\theta')}{\pi(D|X)q(\theta, \theta')\pi(\theta)}\right), \qquad (1)$$

otherwise set $z_{t+1} = z_t$.

Sequential GABC algorithms

Three sequential ABC algorithms have been proposed (Sisson *et al.* (2007), Beaumont *et al.* (2009), Toni *et al.* (2008)) - all of which can be seen to be a special case of the sequential GABC algorithm.

Specify a sequence of target distributions

$$\pi_n(\theta, x) = \frac{\pi_n(D|x)\pi(x|\theta)\pi(\theta)}{C_n} = \frac{\gamma_n(\theta, x)}{C_n}$$

where $\pi_n(D|x)$ has decreasing variance (corresponding to decreasing tolerance δ in uniform SMC-ABC).

Sequential GABC algorithms

Three sequential ABC algorithms have been proposed (Sisson *et al.* (2007), Beaumont *et al.* (2009), Toni *et al.* (2008)) - all of which can be seen to be a special case of the sequential GABC algorithm.

Specify a sequence of target distributions

$$\pi_n(\theta, x) = \frac{\pi_n(D|x)\pi(x|\theta)\pi(\theta)}{C_n} = \frac{\gamma_n(\theta, x)}{C_n}$$

where $\pi_n(D|x)$ has decreasing variance (corresponding to decreasing tolerance δ in uniform SMC-ABC).

At each stage n, we aim to construct a weighted sample of particles that approximates $\pi_n(\theta, x)$.

$$\left\{\left(z_n^{(i)}, W_n^{(i)}\right)\right\}_{i=1}^{N} \text{ such that } \pi_n(z) \approx \sum W_n^{(i)} \delta_{z_n^{(i)}}(\mathrm{dz})$$

where $z_n^{(i)} = (\theta_n^{(i)}, x_n^{(i)})$.

Sequential Monte Carlo (SMC)

If at stage n we use proposal distribution $\eta_n(z)$ for the particles, then we create the weighted sample as follows:

Generic Sequential Monte Carlo - stage n

(i) For $i = 1, \ldots, N$

$$Z_n^{(i)} \sim \eta_n(z)$$

and correct between η_n and π_n

$$w_n(Z_n^{(i)}) = \frac{\gamma_n(Z_n^{(i)})}{\eta_n(Z_n^{(i)})}$$

(ii) Normalize to find weights $\{W_n^{(i)}\}$.

(iii) If effective sample size (ESS) is less than some threshold T, resample the particles and set $W_n^{(i)} = 1/N$. Set $n = n + 1$.

Q: How do we build a sequence of proposals η_n?

Del Moral *et al.* SMC algorithm

We can build the proposal distribution $\eta_n(z)$, from the particles available at time $n-1$.

One way to do this is to propose new particles by passing the old particles through a Markov kernel $K_n(z, z')$.

- For $i = 1, \ldots, N$

$$z_n^{(i)} \sim K_n(z_{n-1}^{(i)}, \cdot)$$

This makes $\eta_n(z) = \int \eta_{n-1}(z')K_n(z', z)\mathrm{dz}'$ – which is unknown in general.

Del Moral *et al.* showed how to avoid this problem by introducing a sequence of backward kernels, L_{n-1}.

Del Moral *et al.* SMC algorithm - step n

(i) Propagate: Extend the particle paths using Markov kernel K_n.

$$\text{For } i = 1, \ldots, N, \qquad Z_n^{(i)} \sim K_n(z_{n-1}^{(i)}, \cdot)$$

(ii) Weight: Correct between $\eta_n(z_{0:n})$ and $\tilde{\pi}_n(z_{0:n})$. For $i = 1, \ldots, N$

$$w_n(z_{0:n}^{(i)}) = \frac{\tilde{\gamma}_n(z_{0:n}^{(i)})}{\eta_n(z_{0:n}^{(i)})} \tag{2}$$

$$= W_{n-1}(z_{0:n-1}^{(i)}) \tilde{w}_n(z_{n-1}^{(i)}, z_n^{(i)}) \tag{3}$$

where

$$\tilde{w}_n(z_{n-1}^{(i)}, z_n^{(i)}) = \frac{\gamma_n(z_n^{(i)}) L_{n-1}(z_n^{(i)}, z_{n-1}^{(i)})}{\gamma_{n-1}(z_{n-1}^{(i)}) K_n(z_{n-1}^{(i)}, z_n^{(i)})} \tag{4}$$

is the incremental weight.

(iii) Normalise the weights to obtain $\{W_n^{(i)}\}$.

(iv) Resample if ESS$< T$ and set $W_n^{(i)} = 1/N$ for all i. Set $n = n + 1$.

SMC with partial rejection control (PRC)

We can add in the rejection control idea of Liu

Del Moral SMC algorithm with Partial Rejection Control - step n

(i) For $i = 1, \ldots, N$

(a) Sample z^* from $\{z_{n-1}^{(i)}\}$ according to weights $W_{n-1}^{(i)}$.

(b) Perturb:

$$z^{**} \sim K_n(z^*, \cdot)$$

(c) Weight

$$w^* = \frac{\gamma_n(z_n^{(i)}) L_{n-1}(z_n^{(i)}, z_{n-1}^{(i)})}{\gamma_{n-1}(z_{n-1}^{(i)}) K_n(z_{n-1}^{(i)}, z_n^{(i)})}$$

(d) PRC: Accept z^* with probability $\min(1, \frac{w^*}{c_n})$. If accepted set $z_n^{(i)} = z^{**}$ and set $w_n^{(i)} = \max(w^*, c_n)$. Otherwise return to (a).

(ii) Normalise the weights to get $W_n^{(i)}$.

GABC versions of SMC

We need to choose

- Sequence of targets π_n
- Forward perturbation kernels K_n
- Backward kernels L_n
- Thresholds c_i.

Del Moral *et al.* showed that the optimum choice for the backward kernels is

$$L_{k-1}^{opt}(z_k, z_{k-1}) = \frac{\eta_{k-1}(z_{k-1})K_k(z_{k-1}, z_k)}{\eta_k(z_k)}$$

This isn't available, but the choice should be made to approximate L^{opt}.

Uniform SMC-ABC

By making particular choices for these quantities we can recover all previously published sequential ABC samplers. For example,

- let π_n be the uniform ABC target using δ_n,

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

- let $K_n(z, z') = K_n(\theta, \theta')\pi(x'|\theta)$
- let $c_1 = 1$ and $c_n = 0$ for $n \geq 2$
- let

$$L_{n-1}(z_n, z_{n-1}) = \frac{\pi_{n-1}(z_{n-1})K_n(z_{n-1}, z_n)}{\pi_{n-1}K_n(z_n)}$$

and approximate $\pi_{n-1}K_n(z) = \int \pi_{n-1}(z')K_n(z', z)\mathrm{dz}'$ by

$$\pi_{n-1}K_n(z) \approx \sum_j W_{n-1}^{(j)} K_n(z_{n-1}^{(j)}, z)$$

then the algorithm reduces to Beaumont *et al.* We recover the Sisson errata algorithm if we add in a further (unnecessary) resampling step. Toni *et al.* is recovered by including a compulsory resampling step.

SMC-GABC

The use of generalised acceptance kernels (rather than uniform) opens up several new possibilies. The direct generalised analogue of previous uniform SMC algorithms is

SMC-GABC

(i) For $i = 1, \ldots, N$

(a) Sample θ^* from $\{\theta_{n-1}^{(i)}\}$ according to weights $W_{n-1}^{(i)}$.

(b) Perturb:

$$\begin{aligned} \theta^{**} &\sim K_n(\theta^*, \cdot) \\ x^{**} &\sim \pi(x|\theta^{**}) \\ w^* &= \frac{\pi_n(D|x^{**})\pi(\theta^{**})}{\sum_j W_{n-1}^{(j)} K_n(\theta_{n-1}^{(j)}, \theta^{**})} \end{aligned} \tag{5}$$

(c) PRC: Accept (θ^{**}, x^{**}) with probability $\min(1, \frac{w^*}{c_n})$. If accepted set $z_n^{(i)} = (\theta^{**}, x^{**})$ and set $w_n^{(i)} = \max(w^*, c_n)$. Otherwise return to (a).

(ii) Normalise the weights to get $W_n^{(i)}$.

SMC-GABC

Note that unlike in uniform ABC, using partial rejection control isn't necessary (the number of particles in uniform ABC would decrease in each step). Without PRC we would need to resample manually as before, according to some criteria (ESS$< T$ say).

Note also that we could modify this algorithm to keep sampling until the effective sample size of the new population is at least as large as some threshold value, N say.

Other sequential GABC algorithms

This is only one particular form of sequential GABC algorithm which arises as a consequence of using

$$L_{n-1}(z_n, z_{n-1}) = \frac{\pi_{n-1}(z_{n-1})K_n(z_{n-1}, z_n)}{\pi_{n-1}K_n(z_n)}$$

If we use a π_n invariant Metropolis-Hastings kernel K_n and let

$$L_{n-1}(z_n, z_{n-1}) = \frac{\pi_n(z_{n-1})K_n(z_{n-1}, z_n)}{\pi_n(z_n)}$$

then we get a new algorithm - a GABC Resample-Move (?) algorithm.

Approximate Resample-Move (with PRC)

RM-GABC

(i) While $ESS < N$

(a) Sample $z^* = (\theta^*, X^*)$ from $\{z_{n-1}^{(i)}\}$ according to weights $W_{n-1}^{(i)}$.

(b) Weight:

$$w^* = \tilde{w}_n(X^*) = \frac{\pi_n(D|X^*)}{\pi_{n-1}(D|X^*)}$$

(c) PRC: With probability $\min(1, \frac{w^*}{c_n})$, sample

$$z_n^{(i)} \sim K_n(z^*, \cdot)$$

where K_n is an MCMC kernel with invariant distribution π_n. Set $i = i + 1$.
Otherwise, return to (i)(a).

(ii) Normalise the weights to get $W_n^{(i)}$. Set $n = n + 1$

Note that because the incremental weights are independent of z_n we are able to swap the perturbation and PRC steps.

Approximate RM

This algorithm is only likely to work well when $\pi_n \approx \pi_{n-1}$

For ABC type algorithms we can make sure this is the case by reducing the variance of $\pi_n(D|X)$ slowly.

Notice that because the algorithm weights the particles with the new weight before deciding what to propogate forwards, we can potentially save on the number of simulator evaluations that are required.

Another advantage is that the weight is of a much simpler form, whereas previously we had an $O(N^2)$ operation at every iteration

$$w^* = \frac{\pi_n(D|x^{**})\pi(\theta^{**})}{\sum_j W_{n-1}^{(j)} K_n(\theta_{n-1}^{(j)}, \theta^{**})}$$

(this is unlikely to be a concern unless the simulator is very quick).

A potential disadvantage is that initial simulation studies have shown the RM algorithm to be more prone to degeneracy than the other SMC algorithm.

A quick note on summaries

ABC algorithms often include the use of summary statistics, $S(\mathcal{D})$.

Approximate Rejection Algorithm With Summaries

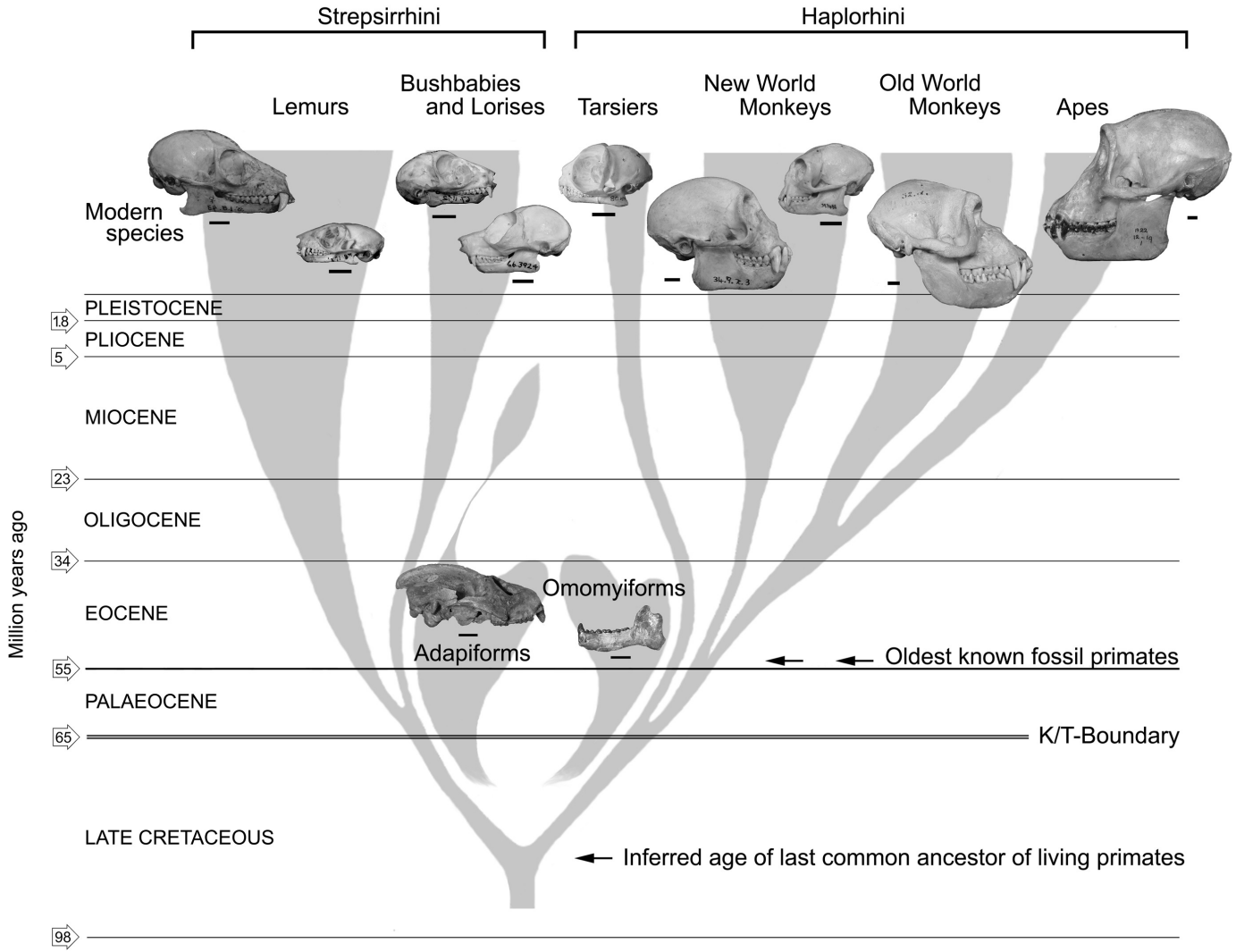
- Draw θ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\rho(S(\mathcal{D}), S(X)) < \delta$

Considerable research effort has focused on automated methods to choose good summaries (sufficiency is not typically achievable) - great if X is some fairly homogenous field of output which we expect the model to reproduce well. Less useful if X is a large collection of different quantities.

Instead ask, what aspects of the data do we expect our model to be able to reproduce? And with what degree of accuracy? $S(D)$ may be highly informative about θ, but if the model was not built to reproduce $S(D)$ then why should we calibrate to it?

Example: Estimating the Primate Divergence

Geologic time



Reconciling molecular and fossil records?

Molecules vs morphology

- Genetic estimates of the primate divergence time are approximately 80-100 mya:

The date has consequences for human-chimp divergence, primate and dinosaur coexistence etc.

Reconciling molecular and fossil records?

Molecules vs morphology

- Genetic estimates of the primate divergence time are approximately 80-100 mya:
  - Uses dna from extant primates, along with the concept of a molecular clock, to estimate the time needed for the genetic diversification.
  - Calibrating the molecular clock relies on other fossil evidence to date other nodes in the mammalian tree.
  - Dates the time of geographic separation

The date has consequences for human-chimp divergence, primate and dinosaur coexistence etc.

Reconciling molecular and fossil records?

Molecules vs morphology

- Genetic estimates of the primate divergence time are approximately 80-100 mya:
  - Uses dna from extant primates, along with the concept of a molecular clock, to estimate the time needed for the genetic diversification.
  - Calibrating the molecular clock relies on other fossil evidence to date other nodes in the mammalian tree.
  - Dates the time of geographic separation
- A direct reading of the fossil record suggests a primate divergence time of 60-65 mya:

The date has consequences for human-chimp divergence, primate and dinosaur coexistence etc.

Reconciling molecular and fossil records?

Molecules vs morphology

- Genetic estimates of the primate divergence time are approximately 80-100 mya:
  - Uses dna from extant primates, along with the concept of a molecular clock, to estimate the time needed for the genetic diversification.
  - Calibrating the molecular clock relies on other fossil evidence to date other nodes in the mammalian tree.
  - Dates the time of geographic separation
- A direct reading of the fossil record suggests a primate divergence time of 60-65 mya:
  - The fossil record, especially for primates, is poor.
  - Fossil evidence can only provide a lower bound on the age.
  - Dates the appearance of morphological differences.

The date has consequences for human-chimp divergence, primate and dinosaur coexistence etc.

Reconciling molecular and fossil records?

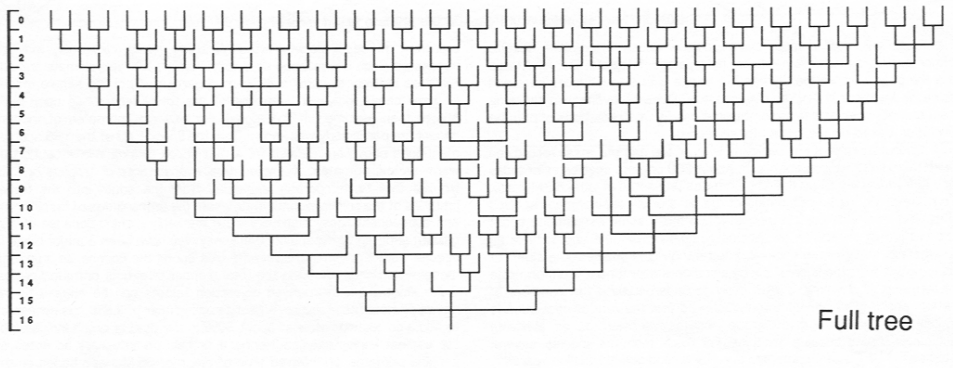
Molecules vs morphology

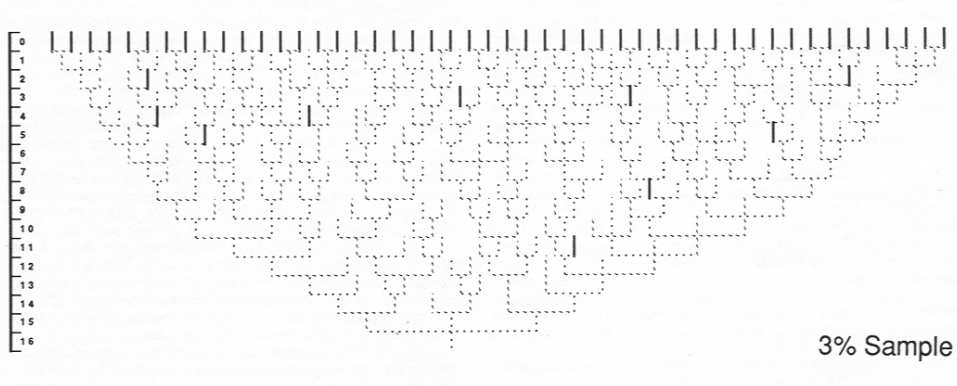
- Genetic estimates of the primate divergence time are approximately 80-100 mya:
  - Uses dna from extant primates, along with the concept of a molecular clock, to estimate the time needed for the genetic diversification.
  - Calibrating the molecular clock relies on other fossil evidence to date other nodes in the mammalian tree.
  - Dates the time of geographic separation
- A direct reading of the fossil record suggests a primate divergence time of 60-65 mya:
  - The fossil record, especially for primates, is poor.
  - Fossil evidence can only provide a lower bound on the age.
  - Dates the appearance of morphological differences.
  - Prevailing view: the first appearance of a species in the fossil record is "... accepted as more nearly objective and basic than opinions as to the time when the group really originated", Simpson, 1965.
  - Oldest primate fossil is 55 million years old.

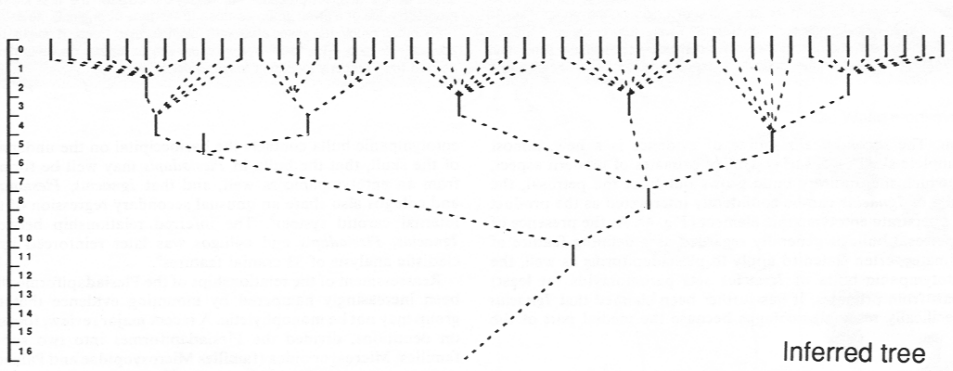
The date has consequences for human-chimp divergence, primate and dinosaur coexistence etc.

Why is this difficult?

Non-repeatable event







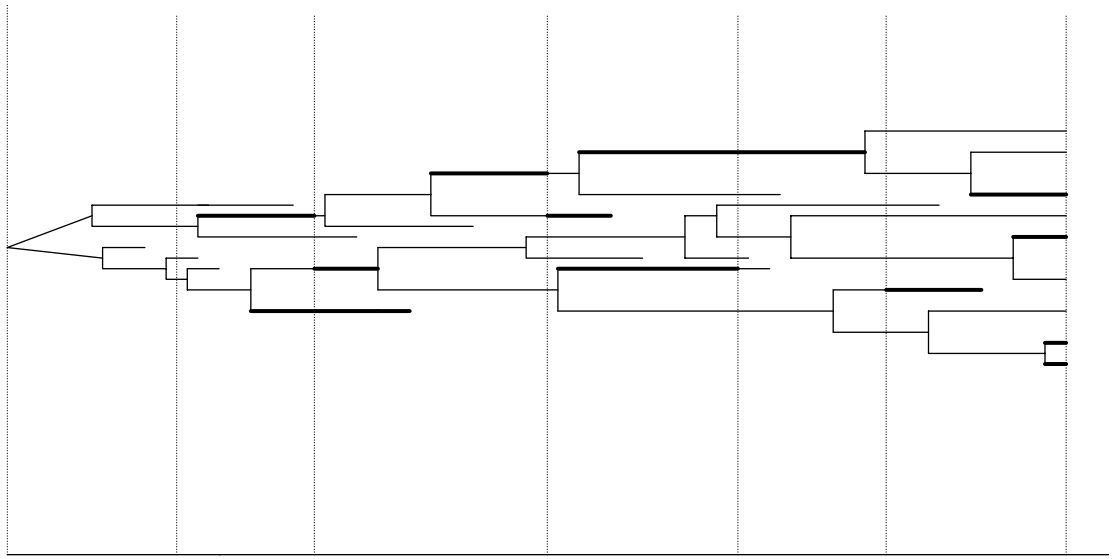
Data

Robert Martin (Chicago) and Christophe Soligo (UCL)

Epoch	k	Time at base of Interval k	Primate fossil counts (D_k)	Anthropoid fossil counts (S_k)
Extant	0	0	376	281
Late-Pleistocene	1	0.15	22	22
Middle-Pleistocene	2	0.9	28	28
Early-Pleistocene	3	1.8	30	30
Late-Pliocene	4	3.6	43	40
Early-Pliocene	5	5.3	12	11
Late-Miocene	6	11.2	38	34
Middle-Miocene	7	16.4	46	43
Early-Miocene	8	23.8	34	28
Late-Oligocene	9	28.5	3	2
Early-Oligocene	10	33.7	22	6
Late-Eocene	11	37.0	30	2
Middle-Eocene	12	49.0	119	0
Early-Eocene	13	54.8	65	
Pre-Eocene	14		0	

- The oldest primate fossil is 54.8 million years old.
- The oldest anthropoid fossil is 37 million years old.

Speciation



An inhomogeneous binary Markov branching process used to model evolution:

- Assume each species lives for a random period of time $\sigma \sim \text{Exponential}(\lambda)$
- Specify the offspring distribution; if a species dies at time t replace it by L_t new species where $\mathbb{P}(L_t = 0) = p_0(t), \mathbb{P}(L_t = 2) = p_2(t)$.

Offspring distribution

If a species dies at time t replace it by L_t new species where $\mathbb{P}(L_t = 0) = p_0(t), \mathbb{P}(L_t = 2) = p_2(t)$.

- Determine the offspring probabilities by fixing the expected population growth $\mathbb{E}(Z(t)) = f(t; \boldsymbol{\lambda})$ and using the fact that

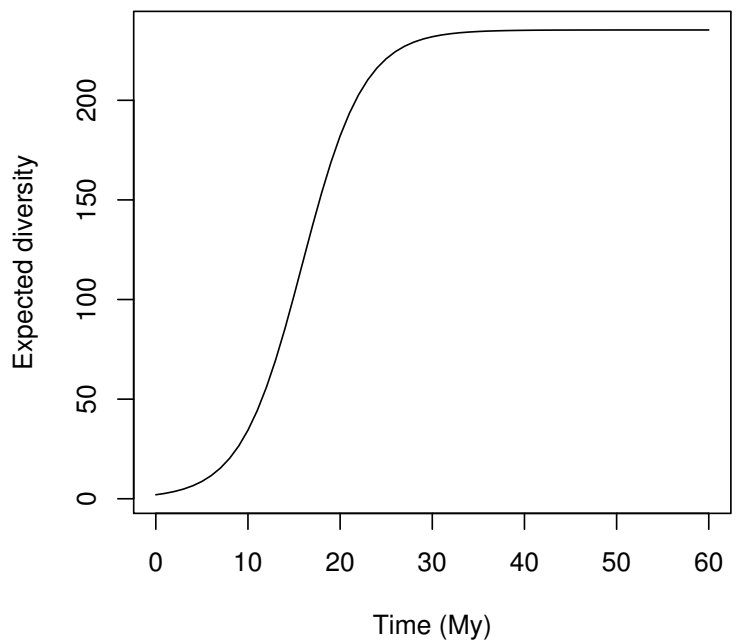
$$\mathbb{E}(Z(t) = n | Z(0) = 2) = 2 \exp\left(\lambda \int_0^t (m(u) - 1) du\right)$$

  where $m(u) = \mathbb{E}L_u$.

For example, assume logistic growth and set

$$\mathbb{E}Z(t) = \frac{2}{\gamma + (1 - \gamma)\exp(-\rho t)}$$

Treat γ and ρ as unknown parameters and infer them in the subsequent analysis.



Fossil Find Model

Recall that time is split into geologic epochs. We have two different models for the number of fossils found in each epoch $\{D_i\}$, given an evolutionary tree $\mathcal{T}$.

Fossil Find Model

Recall that time is split into geologic epochs. We have two different models for the number of fossils found in each epoch $\{D_i\}$, given an evolutionary tree $\mathcal{T}$.

- Binomial Model: each species that is extant for any time in epoch i has a probability α_i of being preserved as a fossil. So that

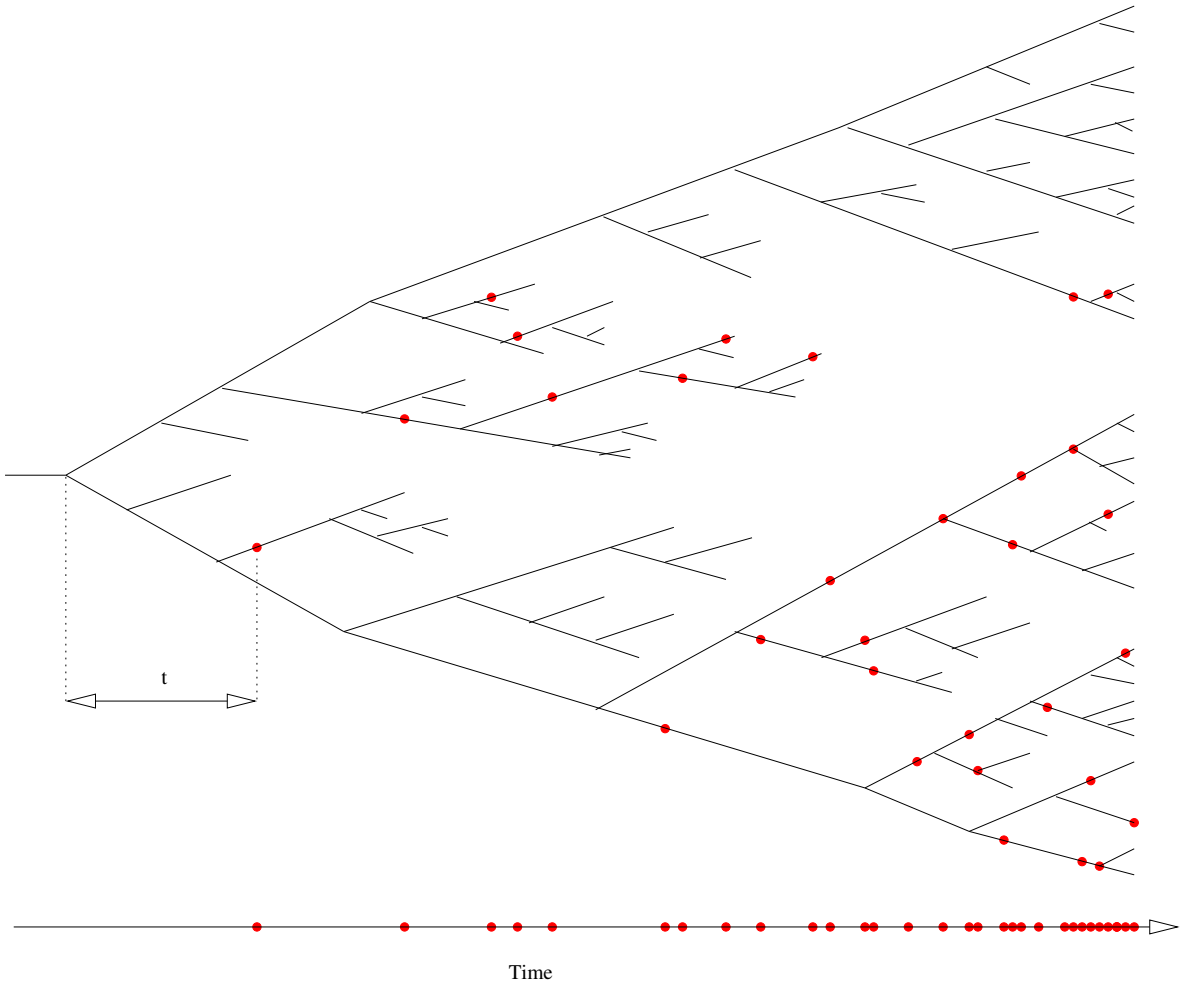
$$\mathbb{P}(D_i|\mathcal{T}) = Bin(N_i, \alpha_i)$$

where $N_i =$ no. species alive during epoch i

Specify the divergence time

Assume

- the primates diverged $54.8 + \tau$ million years ago.



Prior Distributions

We give all parameters prior distributions:

- Temporal gaps between the oldest fossil and the root of the primate and anthropoid trees $\tau \sim U[0, 100]$ and $\tau^* \sim U[0, 100]$.
- Expected life duration of each species $1/\lambda \sim U[2, 3]$
- Growth parameters $\gamma \sim [0.005, 0.015]$ and $\rho \sim U[0, 0.5]$.
- Sampling fractions $\alpha_i \sim U[0, 1]$ (or sampling rates $\beta_i \sim \Gamma(a, b)$).

The aim is to find the posterior distribution of the parameters given the data $\mathcal{D}$, namely $\mathbb{P}(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)$.

Prior Distributions

We give all parameters prior distributions:

- Temporal gaps between the oldest fossil and the root of the primate and anthropoid trees $\tau \sim U[0, 100]$ and $\tau^* \sim U[0, 100]$.
- Expected life duration of each species $1/\lambda \sim U[2, 3]$
- Growth parameters $\gamma \sim [0.005, 0.015]$ and $\rho \sim U[0, 0.5]$.
- Sampling fractions $\alpha_i \sim U[0, 1]$ (or sampling rates $\beta_i \sim \Gamma(a, b)$).

The aim is to find the posterior distribution of the parameters given the data $\mathcal{D}$, namely $\mathbb{P}(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)$.

The likelihood function $\mathbb{P}(\mathcal{D}|\theta)$ is intractable.

$\Downarrow$

MCMC, IS, etc, not possible!
So we use ABC instead.

Choice of metric

We started by using

$$\rho(\mathcal{D}, X) = \sum_{i=0}^{14} (D_i - X_i)^2$$

- This is equivalent to assuming uniform error on a ball of radius $\sqrt{\delta}$ about $\mathcal{D}$.
- It also assumes that errors on each D_i are dependent in some non-trivial manner.
- The error on each D_i is assumed to have the same variance.

Choice of metric

We could move to assuming independent errors by accepting only if

$$(D_i - X_i)^2 \leq \delta_i \text{ for all } i$$

which is equivalent to using the acceptance probability

$$\prod \mathbb{I}_{(D_i - X_i)^2 \leq \delta_i}$$

which we can interpret to be that the error on D_i is uniformly distributed on $[\sqrt{\delta_i}, \sqrt{\delta_i}]$, independently of other errors.

In general, when using summaries $S_1, S_2, \ldots$, it has been suggested that we should choose summaries to be *a priori* independent to increase speed of computation. This will only help if our metric/acceptance kernel assumes independent errors on each S_i.

Uncertainty in the data

The number of extant primates is uncertain:

- Martin (1993) listed 235 primate species

Uncertainty in the data

The number of extant primates is uncertain:

- Martin (1993) listed 235 primate species
- Groves (2005) listed 376 primate species

Uncertainty in the data

The number of extant primates is uncertain:

- Martin (1993) listed 235 primate species
- Groves (2005) listed 376 primate species
- Wikipedia yesterday listed 424 species including
  - the GoldenPalace.com monkey
  - the Avahi cleesei lemur.

Uncertainty in the data

The number of extant primates is uncertain:

- Martin (1993) listed 235 primate species
- Groves (2005) listed 376 primate species
- Wikipedia yesterday listed 424 species including
  - the GoldenPalace.com monkey
  - the Avahi cleesei lemur.

On top of this, there is uncertainty regarding

- whether a bone fragment represents a new species, e.g., homo floresiensis (the hobbit man), or a microcephalic human
- whether two bone fragments represent the same species
- which epoch the species should be assigned to.
- ....

None of these potential sources of errors are accounted for in the model - we only model sampling variation.

Uncertainty in the model

Modelling inevitably involves numerous subjective assumptions. Some of these we judge to be less important.

- Binary trees
- Splitting rather than budding
- Memoryless age distribution

Other assumptions are potentially more influential, particularly where features have been ignored.

- Early Eocene warming (the Paleocene-Eocene Thermal Maximum)
- Warming in the mid-miocene
- Small mass-extinction events in the Cenozoic

We assumed logistic growth for the expected diversity, ignoring smaller fluctuations (we did include the K-T crash).

Uncertainty in the model

Modelling inevitably involves numerous subjective assumptions. Some of these we judge to be less important.

- Binary trees
- Splitting rather than budding
- Memoryless age distribution

Other assumptions are potentially more influential, particularly where features have been ignored.

- Early Eocene warming (the Paleocene-Eocene Thermal Maximum)
- Warming in the mid-miocene
- Small mass-extinction events in the Cenozoic

We assumed logistic growth for the expected diversity, ignoring smaller fluctuations (we did include the K-T crash).
How can we use this information?

- Given that we must add additional uncertainty when using ABC, add it on the parts of the data we are most uncertain about.

Choice of metric

We know that the data from some epochs is more reliable:

- Presumably classification and dating errors are more likely in well sampled epochs - any fossil that is possibly a Cretaceous primate is likely to be well studied, so perhaps we are more confident that $D_{14} = 0$ than that $D_7 = 46$.
- Similarly, large D_i presumably have a larger error than small values of D_i.

Similarly, we know the computer model prediction is more unreliable in some epochs.

- We ignored warm periods in the Eocene and Miocene. During these times primates are believed to have moved away from the tropics, perhaps allowing for more speciation (due to additional space and resources).
- The majority of primate fossils come from the UK, US, France and China, despite our belief that primates originated in the Africa and the observation that nearly all extant species live in tropical or subtropical regions.

An improved metric

In theory, we can account for some of these issues by using the generalised ABC algorithm, using an acceptance probability of the form

$$\pi_\epsilon(X|D) = \prod_{i=0}^{14} \pi_i(X_i|D_i)$$

where $\pi_i(X_i|D_i)$ depends on our belief about measurement and model error on D_i. We might judge that the variance of the measurement error is a function of D_i/D_+ (e.g. interval 14 - the Cretaceous - is likely to have smaller classification error).

Similarly, the model ignores several known features in the Cenozoic, such as warming events. Consequently, we could reduce the importance of the prediction for intervals 11-13 (the Eocene) by allowing a larger error variance during these intervals (we could also allow biases).

An improved metric

In practice, it is a difficult elicitation exercise to specify the errors, and to convolve all the different sources of error.

It is also a difficult computational challenge. Two ideas that might help:

- We can use the fact that we know the distribution of D_i given N_i, the number of simulated species, to help break down the problem (removing the sampling process from the simulation). For example, using the acceptance probability

$$\mathbb{P}(accept) \propto \pi(X_i|D_i) = \begin{cases} 1 \text{ if } D_i = X_i \\ 0 \text{ otherwise} \end{cases}$$

  is equivalent to using

$$\mathbb{P}(accept) \propto \binom{N_i}{D_i} \alpha_i^{D_i} (1-\alpha_i)^{N_i - D_i}$$

  and we can use $N_i = D_i/\alpha_i$ to find a normalising constant.
- $\pi_\epsilon(X|D) = \prod_{i=0}^{14} \pi_i(X_i|D_i)$ provides a sequential structure to the problem that might allow particle methods to be used.

Conclusions

Approximate Bayesian Computation gives exact inference for the wrong model.

- To move beyond inference conditioned on a perfect model hypothesis, we should account for model error.
- ABC algorithms can be considered as adding additional variability on to the model outputs.
- We can 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.
- Relating simulators to reality is hard, even with expert knowledge. However, most modellers have beliefs about where their simulator is accurate, and where it is not.
- If done wisely, ABC can be viewed not as an approximate form of Bayesian inference, but instead as coming closer to the inference we want to do.

Conclusions

Approximate Bayesian Computation gives exact inference for the wrong model.

- To move beyond inference conditioned on a perfect model hypothesis, we should account for model error.
- ABC algorithms can be considered as adding additional variability on to the model outputs.
- We can 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.
- Relating simulators to reality is hard, even with expert knowledge. However, most modellers have beliefs about where their simulator is accurate, and where it is not.
- If done wisely, ABC can be viewed not as an approximate form of Bayesian inference, but instead as coming closer to the inference we want to do.

Thank you for listening!