# A modelling approach to ABC

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

r.d.wilkinson@nottingham.ac.uk

Leeds - September 2012

# Talk Plan

1. Introduction - the need for simulation based methods
2. (Generalised) ABC as continuation of the modelling process
   - Uniform ABC
   - Difficulties with ABC
   - Generalised ABC
3. Dating species divergence times
4. Concluding remarks
5. Relationship of ABC to other methods

# The need for simulation based methods

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)'*

# The need for simulation based methods

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)'*

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

# The need for simulation based methods

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)'*

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 $\theta$ 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 $\theta$ that explain this data.

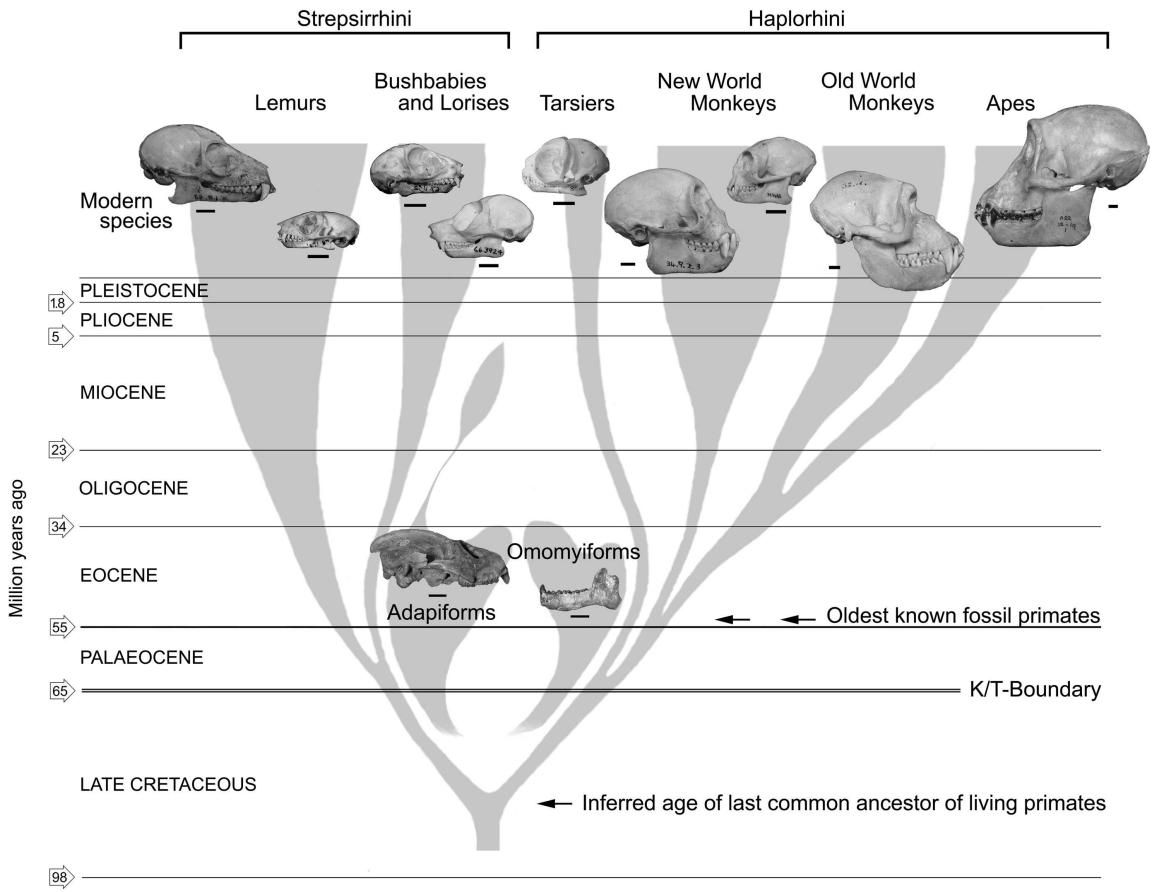
For Bayesians, this is a question of finding the posterior distribution

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

posterior $\propto$

prior $\times$ likelihood

The likelihood isn't be th simulator pdf



# Intractability

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

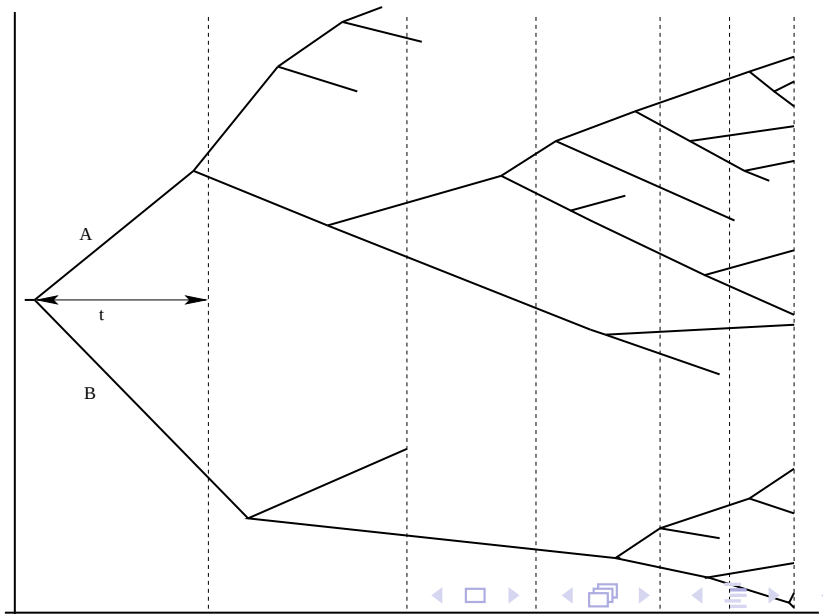
A Bayesian inference problem is intractable if

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

is unknown and can't be evaluated (unknown is subjective). I.e., if the analytic distribution of the simulator, $f(\theta)$, run at $\theta$ is unknown.
Note, this is worse than the usual normalising constant intractability

**Example:**

The density of the cumulative process of a branching process is unknown in general.
We could probably impute everything, but this will be costly.



# Approximate Bayesian Computation (ABC)

# 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 are becoming very popular in the biological sciences.

Heuristic versions of the algorithm exist in most modelling communities.

# 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 are becoming very popular in the biological sciences.

Heuristic versions of the algorithm exist in most modelling communities. ABC methods can be crude but they have an important role to play.

- Scientists are building simulators (intractable ones), and fitting them to data .
  - There is a place for simple methods that can be credibly applied.
  - Likelihood methods for complex simulators are complex.
  - Modelling is something that can be done well by scientists not trained in complex statistical methods.

# Uniform ABC algorithms

## Uniform ABC

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

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

# Uniform ABC algorithms

## Uniform ABC

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

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

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

$\epsilon$ reflects the tension between computability and accuracy.

The hope is that $\pi_{ABC}(\theta) \approx \pi(\theta|D, PSH)$ for $\epsilon$ small, where PSH is perfect simulator hypothesis

# Uniform ABC algorithms

## Uniform ABC

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

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

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

$\epsilon$ reflects the tension between computability and accuracy.

The hope is that $\pi_{ABC}(\theta) \approx \pi(\theta|D, PSH)$ for $\epsilon$ small, where PSH is perfect simulator hypothesis
There are uniform ABC-MCMC, ABC-SMC, ABC-EM, ABC-EP, ABC-MLE algorithms, etc.

# Two ways of thinking

We think about linear regression in two ways

- Algorithmically: find the straight line that minimizes the sum of the squared errors.
- Probability model: we have a linear model with Gaussian errors, and we estimate the parameters using maximum-likelihood.

# Two ways of thinking

We think about linear regression in two ways

- Algorithmically: find the straight line that minimizes the sum of the squared errors.
- Probability model: we have a linear model with Gaussian errors, and we estimate the parameters using maximum-likelihood.

We think about the Kalman filter in two ways:

- Algorithmically: linear quadratic estimation - find the best guess at the trajectory using linear dynamics and a quadratic penalty function
- Probability model: the (Bayesian) solution to the linear Gaussian filtering problem.

# Two ways of thinking

We think about linear regression in two ways

- Algorithmically: find the straight line that minimizes the sum of the squared errors.
- Probability model: we have a linear model with Gaussian errors, and we estimate the parameters using maximum-likelihood.

We think about the Kalman filter in two ways:

- Algorithmically: linear quadratic estimation - find the best guess at the trajectory using linear dynamics and a quadratic penalty function
- Probability model: the (Bayesian) solution to the linear Gaussian filtering problem.

The same dichotomy exists for ABC.

# Algorithmic view of ABC

I'd suggest that most of the early ABC developments have been in the algorithmic tradition.

1. Find a good metric, $\rho$ - e.g., $L_2$ norm
2. Find a good $\epsilon$ - e.g., best 1% of simulations?
3. Find a good summary $S(D)$

The choices made are usually not motivated by modelling considerations.

Poor choices for any of these aspects can have unintended consequences.

# Choice of metric $\rho$

Consider the following system

$$X_{t+1} = X_t + N(0, \sigma^2) \tag{1}$$
$$Y_t = X_t + N(0, \tau^2) \tag{2}$$

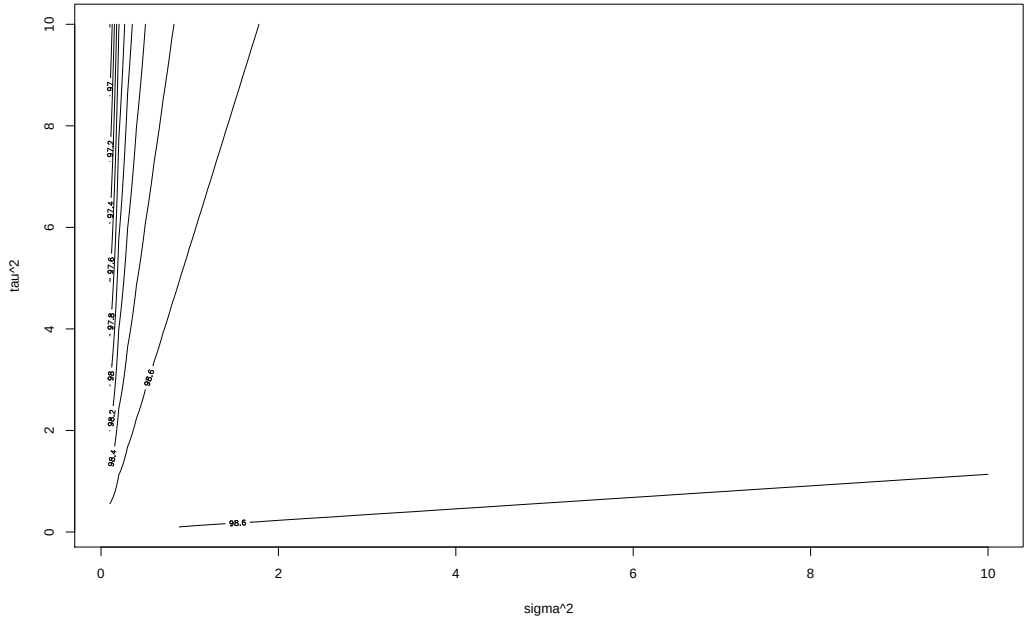
where we want to estimate measurement error $\tau$ and model error $\sigma$. Default choice of metric (or similar)

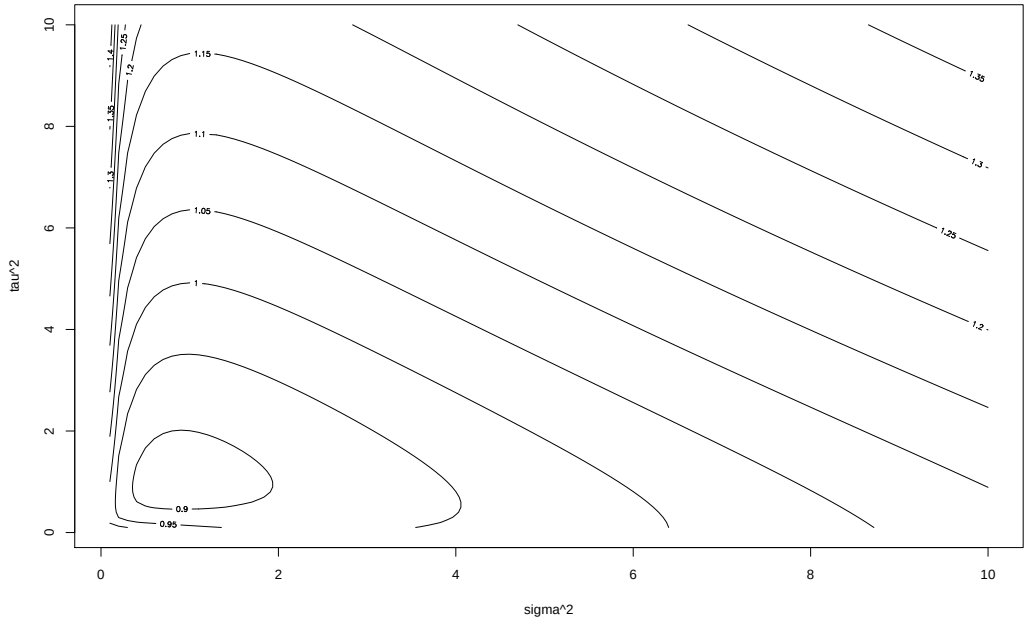
$$\rho(Y, y^{obs}) = \sum (y_t^{obs} - Y_t)^2$$

or instead we could use the CRPS (a proper scoring rule)

$$\rho(y^{obs}, F(\cdot)) = \sum crps(y_t^{obs}, F_t(\cdot)) = \sum_t \int (F_t(u) - \mathbb{I}_{y_t \leq u})^2 \mathrm{du}$$

where $F_t(\cdot)$ is the distribution function of $Y_t | y_{1:t-1}$.


tau^2
sigma^2
0
2
4
6
8
10
97
97.2
97.4
97.6
97.8
98
98.2
98.4
98.6
98.6


tau^2
sigma^2
0
2
4
6
8
10
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35

# Choice of tolerance $\epsilon$

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, known variance $\sigma^2$, $\mu \sim U[a, b]$.
The mean of the data is sufficient for $\mu$ and so we can compare data sets by comparing means. Using $\rho(\mathbf{D}, \mathbf{X}) = |\bar{\mathbf{D}} - \bar{\mathbf{X}}|$ gives the following ABC:

- Pick $\mu$ from the prior distribution,
- Simulate $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$,
- Accept $\mu$ if $|\bar{X} - \bar{D}| \leq \epsilon$.

# Choice of tolerance $\epsilon$

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, known variance $\sigma^2$, $\mu \sim U[a, b]$.
The mean of the data is sufficient for $\mu$ and so we can compare data sets by comparing means. Using $\rho(\mathbf{D}, \mathbf{X}) = |\bar{\mathbf{D}} - \bar{\mathbf{X}}|$ gives the following ABC:

- Pick $\mu$ from the prior distribution,
- Simulate $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$,
- Accept $\mu$ if $|\bar{X} - \bar{D}| \leq \epsilon$.

Calculation of $\pi_{ABC}(\mu)$ and $\pi(\mu|D)$ is possible, and we can show that

$$\mathbb{V}\mathrm{ar}_{ABC}(\mu) \approx \frac{\sigma^2}{n} + \frac{1}{3}\epsilon^2$$

$$d_{TV}(\pi_{ABC}(\mu), \pi(\mu|\mathcal{D})) \approx \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2)$$

The tolerance required for a given accuracy depends on the size of the posterior variance $\sigma^2/n$. For small posterior variances we shall need to use a smaller value of $\epsilon$, whereas for large variances we can use a larger $\epsilon$.

# Choice of tolerance

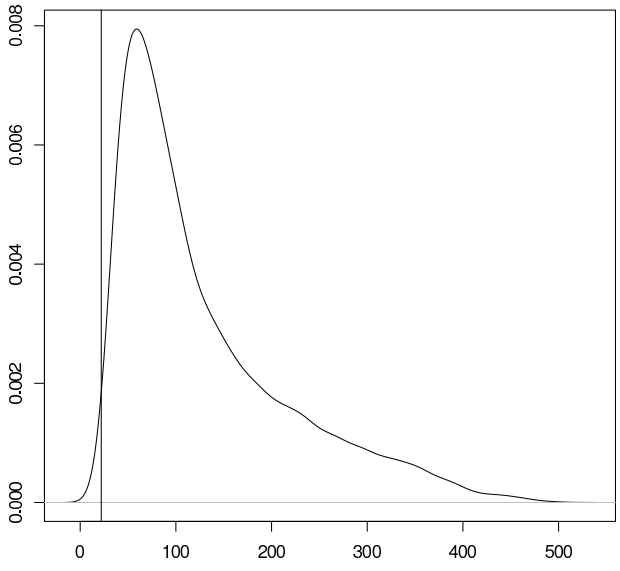
Depending on how the problem is set up, it is possible to use too-small a value for $\epsilon$ (e.g., when taking the best 1%).
e.g., inferring the parameters of a dynamic model with Guassian noise:

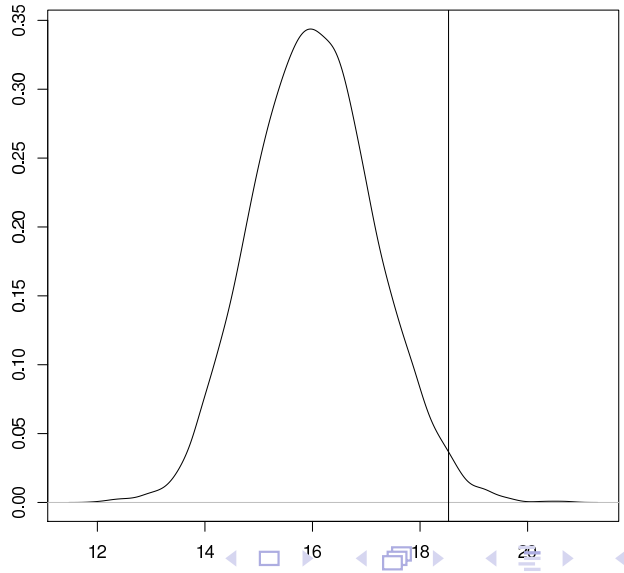
$$X_{1:T} \sim SDE(\theta) \qquad D_t \sim N(X_t, \sigma^2)$$

The default ABC algorithm (Toni *et al.* 2009) is

$\theta \sim \pi(\theta)$; $X_{1:T} \sim SDE(\theta)$ (don't simulate measurement error)
Accept $\theta$ if $\rho(D_{1:T}, X_{1:T}) \leq \epsilon$

$\rho(D_{1:T}, X_{1:T})$ as $X_{1:T} \sim f(\hat{\theta})$ and as $D_{1:T} \sim \pi(D_{1:T}|\hat{X}_{1:T})$



# Choice of summary $S$

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

- Accept $\theta$ 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 $\theta$, but if the model was not built to reproduce $S(D)$ then why should we calibrate to it?

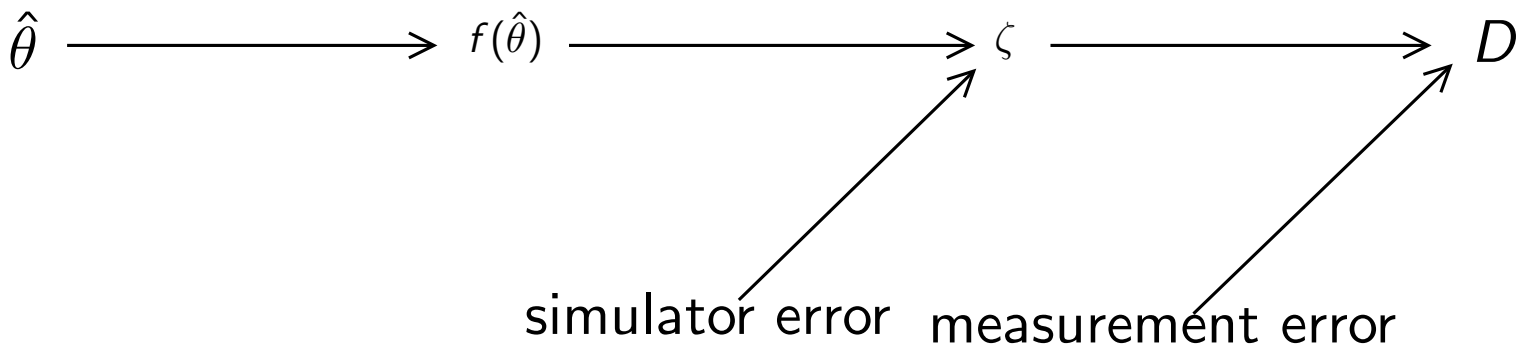
- Cf. Simon Wood's (2010) rejection of phase sensitive summaries in dynamical models

# Calibration framework

Lets now consider the probabilistic interpretation of ABC.

The Bayesian calibration framework from the computer experiment literature:

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




See, for example, Kennedy and O'Hagan (2001) or Goldstein and Rougier (2009).

# 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 - call it the *acceptance kernel*.
- $\pi(x|\theta)$ is the likelihood function of the simulator (ie not relating to reality)

The posterior is

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

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

The posterior is

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

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 $(\theta, x)$ space

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

# How does ABC relate to calibration?

## Wilkinson 2008 and forthcoming

Consider how this relates to ABC:

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

Lets sample from this using the rejection algorithm with instrumental distribution

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

# Generalized ABC (GABC)

Wilkinson 2008, Fearnhead and Prangle 2012

The rejection algorithm then becomes

## Generalized rejection ABC (Rej-GABC)

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

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

# Generalized ABC (GABC)

Wilkinson 2008, Fearnhead and Prangle 2012

The rejection algorithm then becomes

## Generalized rejection ABC (Rej-GABC)

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

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

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 $\theta$ 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 $\theta$ from the uniform ABC algorithm (with $\rho(D, X) = |D - X|$) are samples from the posterior distribution of $\theta$ 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 $\theta$ from the uniform ABC algorithm (with $\rho(D, X) = |D - X|$) are samples from the posterior distribution of $\theta$ 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!

# 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 $(\theta_i, x_i)$ 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 $(\theta_i, x_i)$ weight $w_i = \pi(D|x_i)$.

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

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

This allows us to see part of the Beaumont *et al.* 2002 as using weighted averages with a partial rejection control to estimate posterior integrals

# 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 $(\theta_i, x_i)$ weight $w_i = \pi(D|x_i)$.

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

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

This allows us to see part of the Beaumont *et al.* 2002 as using weighted averages with a partial rejection control to estimate posterior integrals Generalisations of the MCMC-ABC and SMC-ABC algorithms to this framework are available.

# Acceptance kernel $\pi(D|X)$ as an extension of modelling

Using ABC is equivalent to adding additional variability into the model.

- $\exists$ many interesting papers saying how to make this variability small
- Instead ask, given that we are stuck with this additional variability, can we use it in a useful manner, or if not, how can we make sure it does little harm?

# Acceptance kernel $\pi(D|X)$ as an extension of modelling

Using ABC is equivalent to adding additional variability into the model.

- $\exists$ many interesting papers saying how to make this variability small
- Instead ask, given that we are stuck with this additional variability, can we use it in a useful manner, or if not, how can we make sure it does little harm?

How do we relate the simulator to the observations $\pi(D|S)$

- Measurement/sampling error on $D$
  - Measurement error may be built into the simulator. Could we remove it and use the ABC to do this?

# Acceptance kernel $\pi(D|X)$ as an extension of modelling

Using ABC is equivalent to adding additional variability into the model.

- $\exists$ many interesting papers saying how to make this variability small
- Instead ask, given that we are stuck with this additional variability, can we use it in a useful manner, or if not, how can we make sure it does little harm?

How do we relate the simulator to the observations $\pi(D|S)$

- Measurement/sampling error on $D$
  - Measurement error may be built into the simulator. Could we remove it and use the ABC to do this?
- Discrepancy between the simulator and reality
  - In a deterministic model setting, Goldstein and Rougier 2008, and Kennedy and O'Hagan 2001 (amongst others), have offered advice for thinking about model discrepancies.
  - For statistical models, simulator error is a less clear concept.

# Comments

- Prior specification of the distribution of the model error is hard!

# Comments

- Prior specification of the distribution of the model error is hard!
- For many models, the variance of the model and measurement error may not be large enough to make the ABC acceptance rate high enough to generate any successes. This is telling us something has been mis-specified and we need to rethink the choices we have made.

# Comments

- Prior specification of the distribution of the model error is hard!
- For many models, the variance of the model and measurement error may not be large enough to make the ABC acceptance rate high enough to generate any successes. This is telling us something has been mis-specified and we need to rethink the choices we have made.
- It can help us interpret our results, and highlight where we may being doing something undesirable with the modeling or inference.

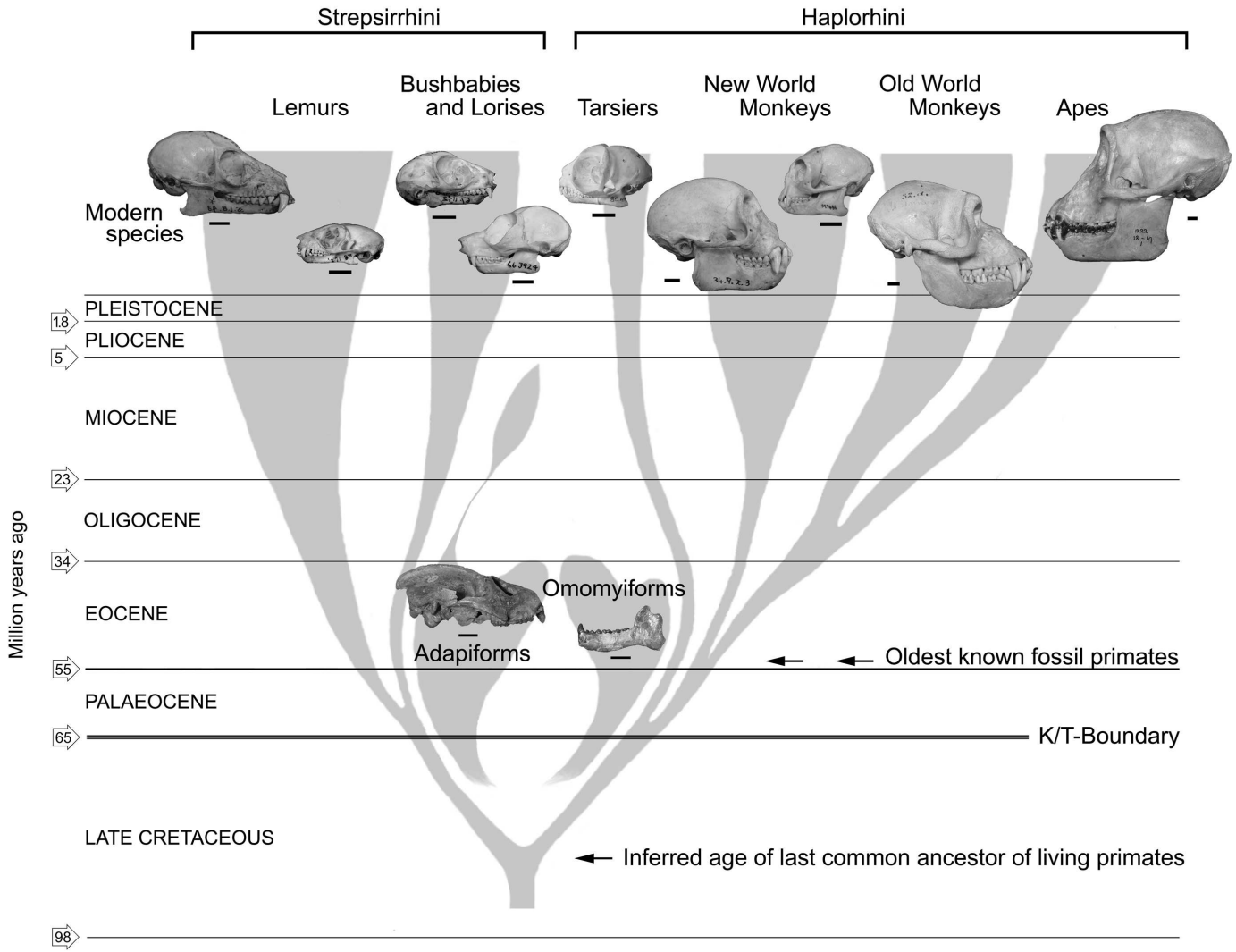
# Comments

- Prior specification of the distribution of the model error is hard!
- For many models, the variance of the model and measurement error may not be large enough to make the ABC acceptance rate high enough to generate any successes. This is telling us something has been mis-specified and we need to rethink the choices we have made.
- It can help us interpret our results, and highlight where we may being doing something undesirable with the modeling or inference.
- It can help guide our choice of metric. We don't expect any model to perfectly fit the data, and ABC can be viewed as adding in enough variability to allow a fit to be found. Given that we are adding in variability, we can control where it is placed.

# Dating Primate Divergences

# Estimating the Primate Divergence Time

## Wilkinson, Steiper, Soligo, Martin, Yang, Tavaré 2011



# 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?

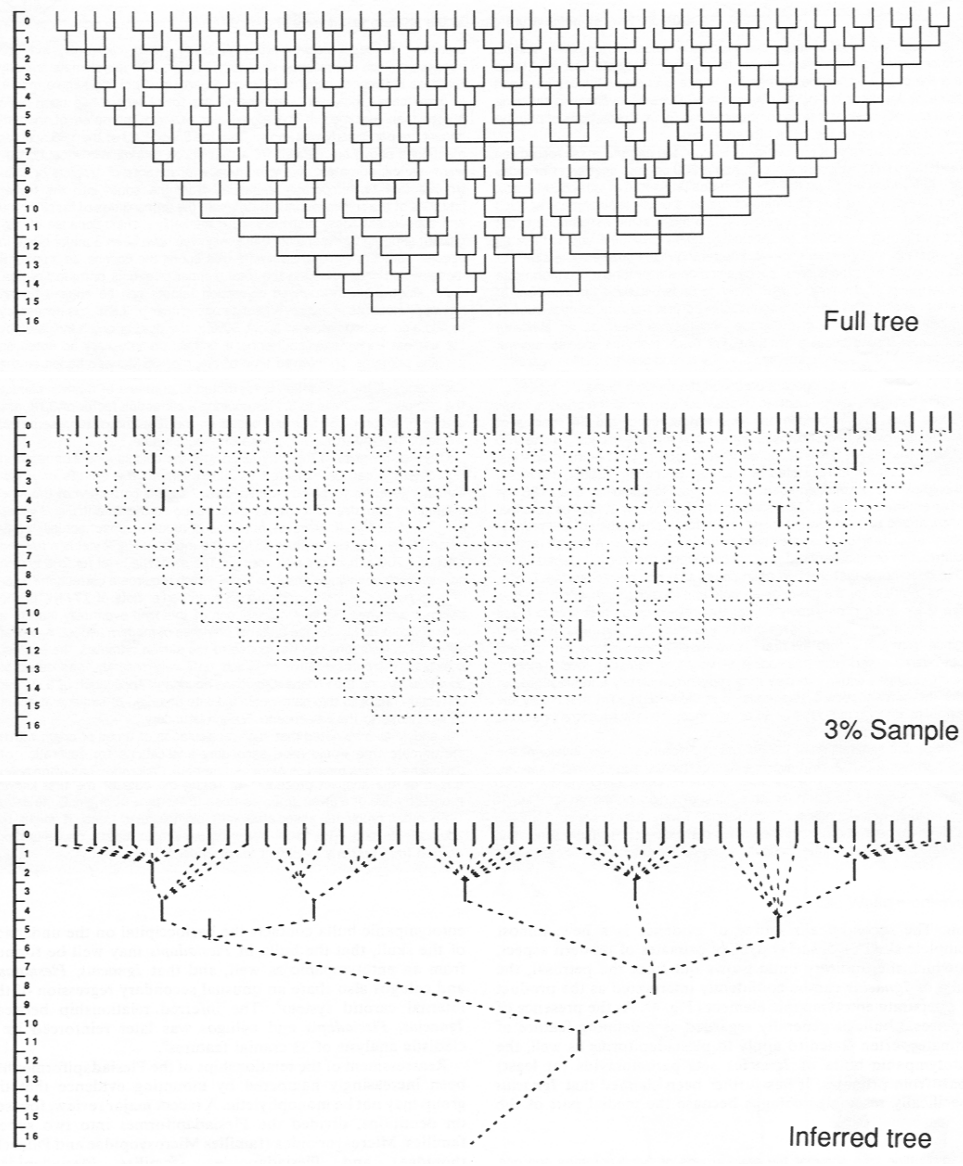
## Molecules vs morphology

- 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



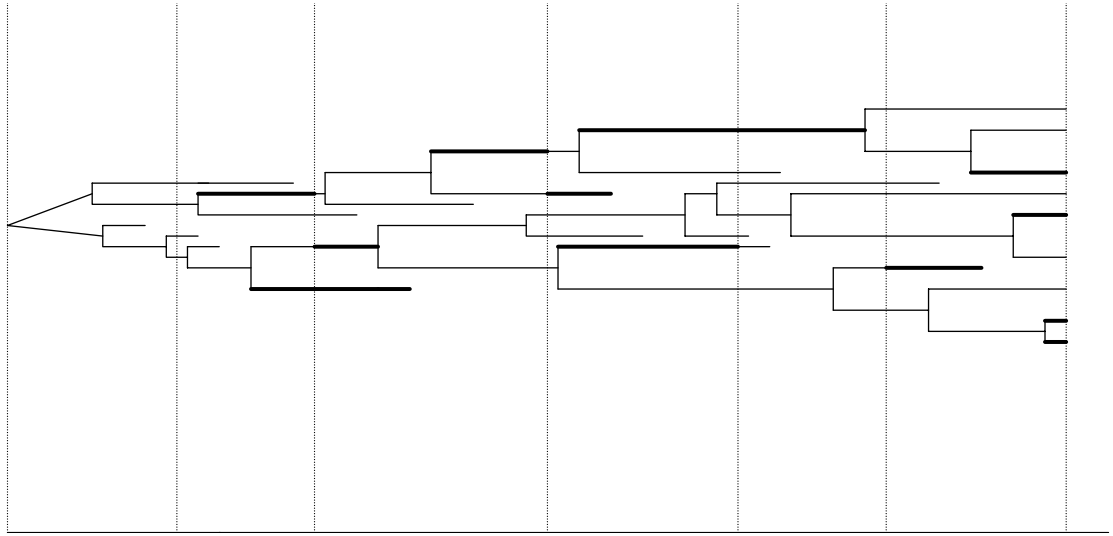
# Data

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 - Dynamical system model



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

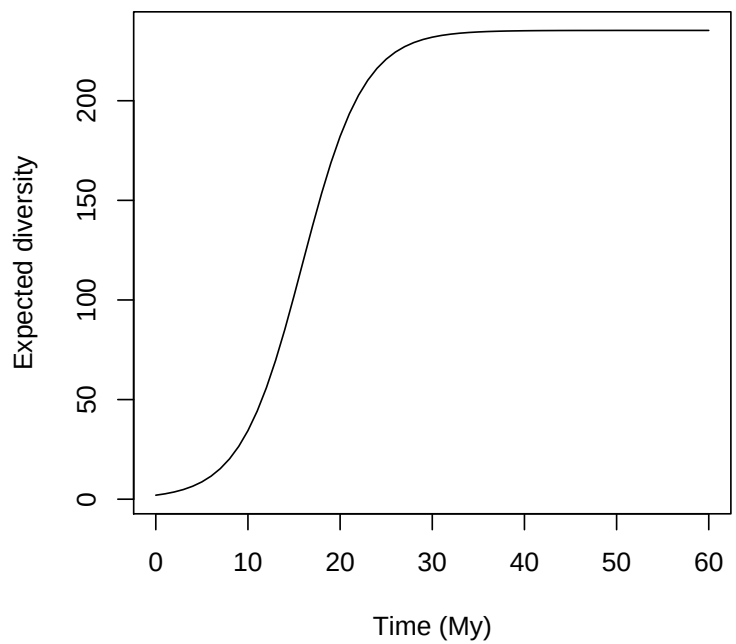
$$\mathbb{E}(Z(t) = n | Z(0) = 2) = 2 \exp\left(\lambda \int_0^t (m(u) - 1) du\right)$$

  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 $\gamma$ and $\rho$ 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 $\mathcal{T}$. The simplest is

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

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

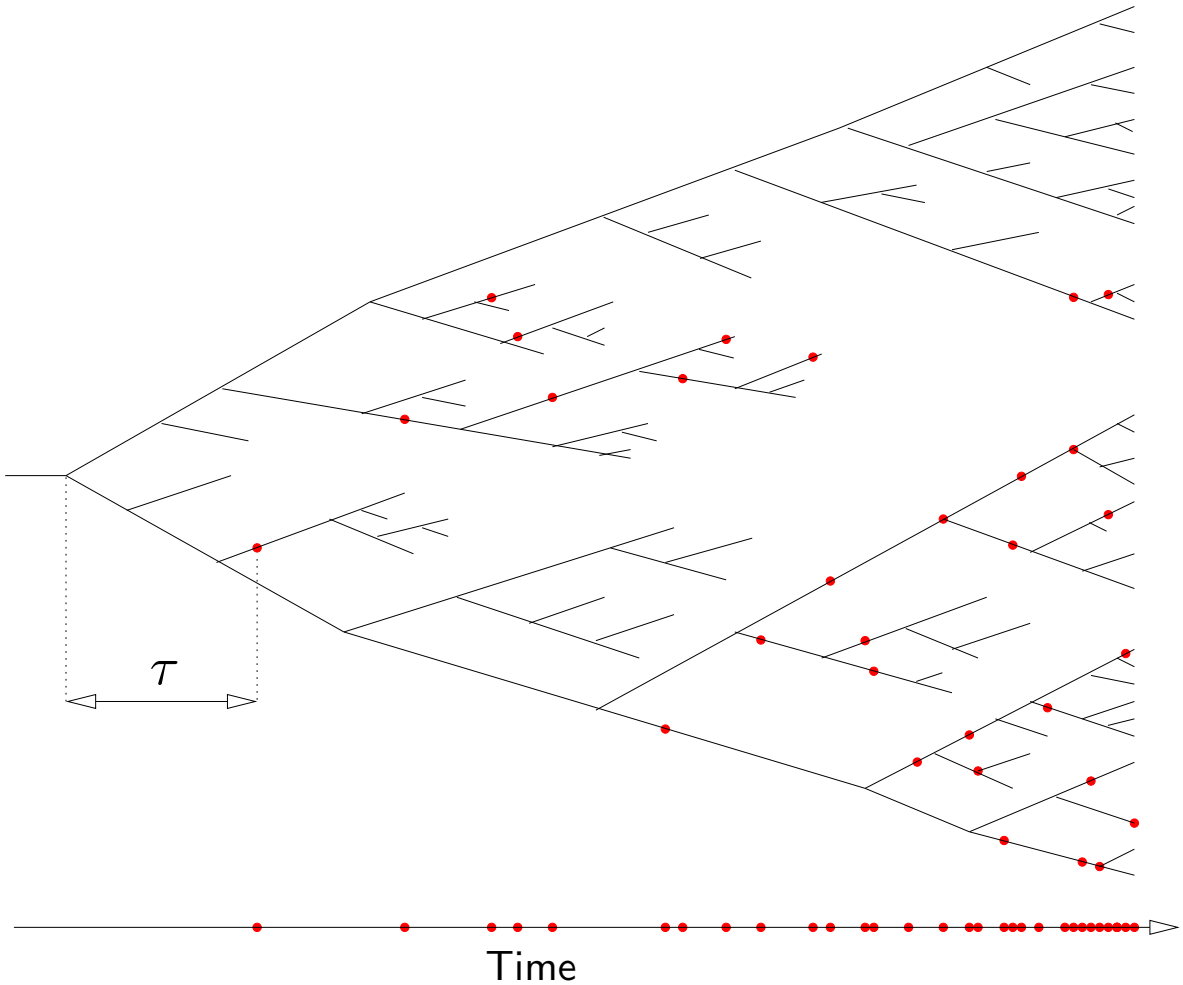
  where $N_i =$ no. species alive during epoch $i$

We can also include known mass extinction events such as the K-T crash.

# 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 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.

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.

# 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 $D_i$.

# 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

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 $D_i$ (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

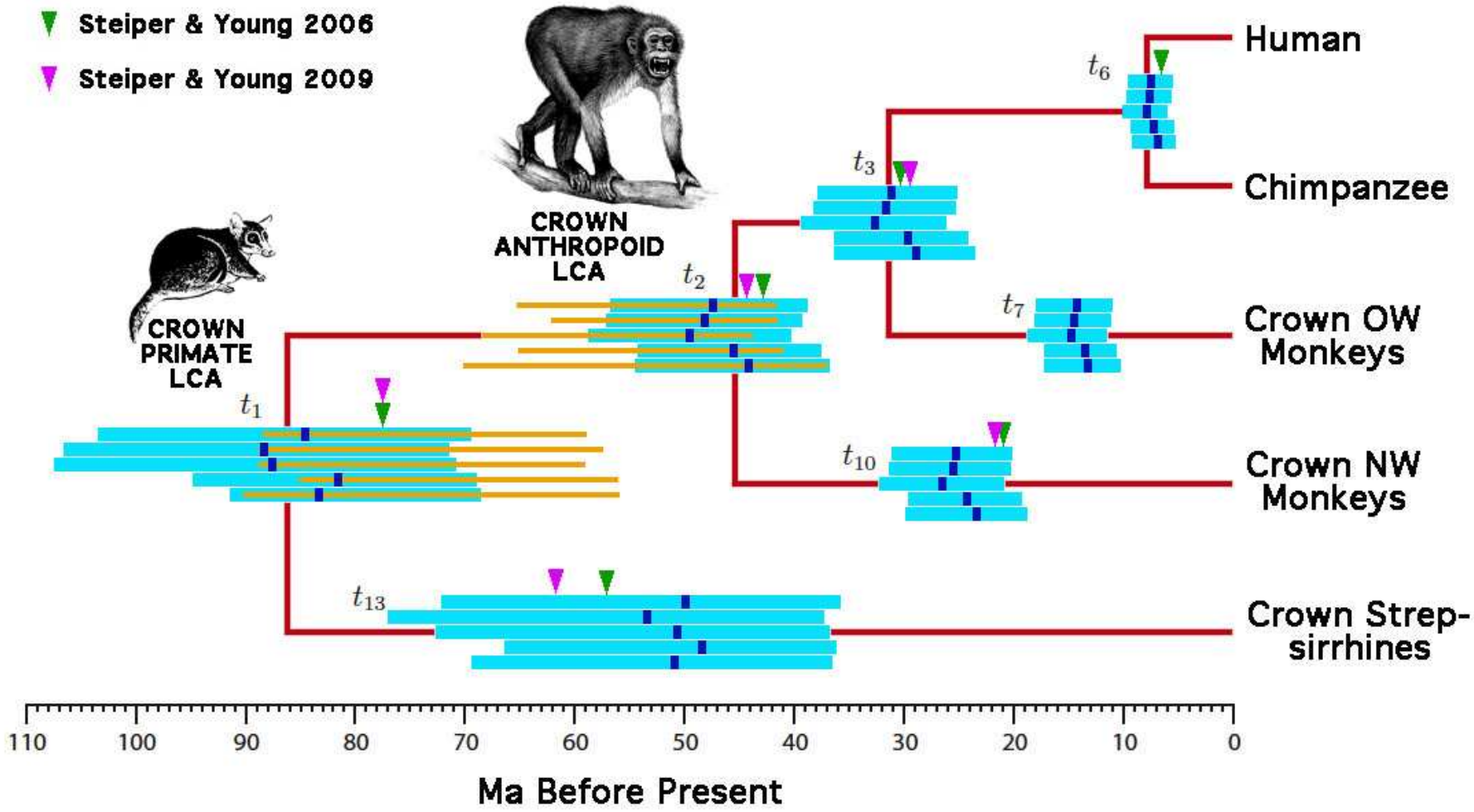
$$\mathbb{P}(accept) \propto \pi(D_i|X_i) = \begin{cases} 1 \text{ if } D_i = X_i \\ 0 \text{ otherwise} \end{cases}$$

  is equivalent to using

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

- $\pi_\epsilon(X|D) = \prod_{i=0}^{14} \pi_i(X_i|D_i)$ provides a sequential structure to the problem that allows particle methods to be used.

# An integrated molecular and palaeontological analysis




The primate fossil record does not precisely constrain the primate divergence time

# Propagating uncertainty forwards

ScienceDaily®

Your source for the latest research news

Did eyes evolve? Look clearly at all the evidence. You can't ignore t

Buy Trilobites in stock Free Authentication Certificate 100s of Kids

The DNA Ancestry Project Discover Your Ancestry with DNA. Find

News | Articles | Videos | Images | Books

Health & Medicine | Mind & Brain | Plants & Animals | Earth & Climate | Space & Time | Matter &

**Science News** Share Blog Cite

## New Statistical Model Moves Human Evolution Back Three Million Years

*ScienceDaily (Nov. 5, 2010)* — Evolutionary divergence of humans and chimpanzees likely occurred some 8 million years ago rather than the 5 million year estimate widely accepted by scientists, a new statistical model suggests.


*A new statistical model suggests that evolutionary divergence of humans from chimpanzees likely occurred some 8 million years ago, rather than the 5 million year estimate widely accepted by scientists. (Credit: iStockphoto/Eric Gevaert)*

**See Also:**

**Plants & Animals**
- Evolutionary Biology
- Nature

**Computers & Math**
- Statistics
- Computer Modeling

**Fossils & Ruins**
- Fossils
- Evolution

**Reference**
- Hominidae
- Multiregional hypothesis

The revised estimate of when the human species parted ways from its closest primate relatives should enable scientists to better interpret the history of human evolution, said Robert D. Martin, curator of biological anthropology at the Field Museum, and a co-author of the new study appearing in the journal *Systematic Biology*.

Working with mathematicians, anthropologists and molecular biologists, Martin has long sought to integrate evolutionary information derived from genetic material in various species with the fossil record to get a more complete picture.

# Conclusions

Scientists are building simulators (intractable ones), and fitting them to data

- There is a place for simple (crude?) inference methods such as ABC
- Likelihood inference is hard, but modelling is something that can be done well by scientists unfamiliar with statistical methods.

# Conclusions

Scientists are building simulators (intractable ones), and fitting them to data

- There is a place for simple (crude?) inference methods such as ABC
- Likelihood inference is hard, but modelling is something that can be done well by scientists unfamiliar with statistical methods.

We can improve the application of ABC algorithms by treating them as modelling approaches.

- view ABC as implicitly defining a probability model for the relationship between the data and the simulator.
- We can generalise ABC algorithms to move beyond the use of uniform error structures - model error etc
  - Relating simulators to reality is hard, even with expert knowledge.
- If done wisely, hopefully 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

Scientists are building simulators (intractable ones), and fitting them to data

- There is a place for simple (crude?) inference methods such as ABC
- Likelihood inference is hard, but modelling is something that can be done well by scientists unfamiliar with statistical methods.

We can improve the application of ABC algorithms by treating them as modelling approaches.

- view ABC as implicitly defining a probability model for the relationship between the data and the simulator.
- We can generalise ABC algorithms to move beyond the use of uniform error structures - model error etc
  - Relating simulators to reality is hard, even with expert knowledge.
- If done wisely, hopefully 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!

# 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.
  - - moves ABC from being an algorithm, to an extension of modelling.
- allows for new ABC algorithms, as (non-trivial) importance sampling algorithms are now possible.
- allows us to interpret the results of ABC
- 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 $\pi(D|X)$
  - - another way to view ABC is as a smoothing of the simulator likelihood

$$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{d}\theta$$

This flattens out the simulator likelihood (cf Wood 2010).

# ABC as likelihood thresholding/history-matching

Consider a deterministic SSM, for example an ode, with observation error

$$x_{t+1} = f(x_t, \theta) \tag{3}$$

$$y_t = x_t + e_t \tag{4}$$

If we use uniform-ABC with

$$\rho(y, X) = \sum (y_t - x_t)^2$$

(as done for example in Toni *et al.* 2009 for the deterministic Lotka-Volterra model) then this is equivalent to assuming a Gaussian distribution for $e_t$, but thresholding the likelihood and only accepting $\theta$ if $\pi(y|x) > h(\epsilon)$.

If the prior for $\theta$ is uniform over some interval, then the uniform-ABC posterior will also be uniform, over the range of values which pass through the likelihood thresholding.

If we use a GABC algorithm, with $\pi(y|X) = \exp(-\rho(y, X)/\sigma^2)$ then we are assuming $e_t \sim N(0, \sigma^2)$ and we get the true posterior under this assumption (not the thresholded version).

# History-matching

The history matching version would go something like

0. Build an emulator of the simulator $f(\theta)$
1. Relate the simulator to the system

$$y = f(\theta) + \epsilon$$

   where $\epsilon$ is our simulator discrepancy
2. Relate the system to the data ($e$ represents measurement error)

$$\mathcal{D} = y + e$$

3. Declare $\theta$ implausible if

$$\| \mathcal{D} - \mathbb{E} f(\theta) \| > 3\sigma$$

   where $\sigma^2$ is the combined variance implied by the emulator, discrepancy and measurement error.

# History-matching

The history matching version would go something like

0. Build an emulator of the simulator $f(\theta)$
1. Relate the simulator to the system

$$y = f(\theta) + \epsilon$$

   where $\epsilon$ is our simulator discrepancy
2. Relate the system to the data ($e$ represents measurement error)

$$\mathcal{D} = y + e$$

3. Declare $\theta$ implausible if

$$\| \mathcal{D} - \mathbb{E}f(\theta) \| > 3\sigma$$

   where $\sigma^2$ is the combined variance implied by the emulator, discrepancy and measurement error.

If $\theta$ is not implausible we don't discard it. The result is a region of space that we can't rule out at this stage of the history-match. Usually we'd go through several stages of history matching,

# Relationship between ABC and history-matching

- Using uniform priors with a deterministic simulator, and using uniform-ABC is similar to doing history matching (ignoring the use of emulators in history-matching and a few other tweeks).
- The different waves of history-matching is analogous to choosing a decreasing tolerance scheme $\epsilon_t$ in sequential-ABC methods.

The result of a history-match may be that there is no not-implausible region of parameter space (unlike the result of a likelihood-based MCMC calculation) - note the very different interpretation to ABC.

# Relationship between ABC and history-matching

- Using uniform priors with a deterministic simulator, and using uniform-ABC is similar to doing history matching (ignoring the use of emulators in history-matching and a few other tweeks).
- The different waves of history-matching is analogous to choosing a decreasing tolerance scheme $\epsilon_t$ in sequential-ABC methods.

The result of a history-match may be that there is no not-implausible region of parameter space (unlike the result of a likelihood-based MCMC calculation) - note the very different interpretation to ABC.

We might find that after quantifying all the uncertainties, we can't find any parameter values that cause the model to fit the data. Similarly with GABC, we might find that no matter how careful we are in specifing the errors, we can't find any region of parameter space that fits.

- suggests we've made a mistake somewhere in the specification of the problem.
- do we want methods that fit the model regardless of the quality of the fit, or that find there are no good fits, making us think harder?

# Why/when/should we include simulator discrepancy??

The importance of including a discrepancy seems to depend on whether our simulator is

- Statistical
- Physical

and also on whether our aim is

- explanation
- prediction.

and on the importance of the situation.

When should we be happy saying that our simulator is imperfect, but we're going to fit it anyway and ignore the problem?

# Relationship between ABC and Simon Wood's approach

One way to view Wood 2010 is as an ABC algorithm, but using $\mu_\theta$ and $\Sigma_\theta$ as the summary of $f(\theta)$, and assuming

$$\pi(D|S) = \exp(-1/2(D - \mu_\theta)^T \Sigma_\theta^{-1}(D - \mu_\theta))$$

A crude IS-GABC algorithm version of Wood 2010 would be

- Pick $\theta \sim \pi(\theta)$
- Simulate $s_1, \ldots, s_n \sim f(\theta)$, calculate $\mu_\theta$ and $\Sigma_\theta$.
- Give $\theta$ weight $w = \phi((D - \mu)^2/\Sigma)$

Wood 2010 could be seen as a MCMC-GABC algorithm.

# Relationship between ABC and Simon Wood's approach

One way to view Wood 2010 is as an ABC algorithm, but using $\mu_\theta$ and $\Sigma_\theta$ as the summary of $f(\theta)$, and assuming

$$\pi(D|S) = \exp(-1/2(D-\mu_\theta)^T \Sigma_\theta^{-1}(D-\mu_\theta))$$

A crude IS-GABC algorithm version of Wood 2010 would be

- Pick $\theta \sim \pi(\theta)$
- Simulate $s_1, \ldots, s_n \sim f(\theta)$, calculate $\mu_\theta$ and $\Sigma_\theta$.
- Give $\theta$ weight $w = \phi((D-\mu)^2/\Sigma)$

Wood 2010 could be seen as a MCMC-GABC algorithm.

This can be seen as accounting for the variability of the model run repeatedly at the same input, and then assuming the distribution is Gaussian. If so, running diagnostics such as QQ-plots does make sense.
Alternatively, we could see it as a way of smoothing the simulator likelihood making inference more tractable.
It is also analogous to building an emulator of a deterministic function in history matching, with the difference that in history matching the uncertainty represents lack of knowledge of the simulator output.

# Noisy-ABC

Fearnhead and Prangle (2012) recently suggested a noisy-ABC algorithm:

## Noisy-ABC

Initialise: Let $D' = D + e$ where $e \sim K(e)$ from some kernel $K(\cdot)$.

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

# Noisy-ABC

Fearnhead and Prangle (2012) recently suggested a noisy-ABC algorithm:

## Noisy-ABC

Initialise: Let $D' = D + e$ where $e \sim K(e)$ from some kernel $K(\cdot)$.

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

In my notation, this replaces the obsevered data $D$, with $D'$ drawn from the acceptance kernel - $D' \sim \pi(D'|D)$

If we believe $\pi(D|X)$ relates the simulator to reality, then noisy-ABC is equivalent to adding another dose of measurement/model error to your data and using that in the inference.

The main argument in favour of noisy-ABC is that it is calibrated, unlike standard ABC.

# Calibration

Calibration is a way of assessing probability statements against some idea of *truth*, a base measure $\mathbb{P}$.

- Truth is usually taken to be reality
- We are well-calibrated if $p\%$ of all predictions reported at probability $p$ are true. Idea goes back at least to Dawid (1984).

# Calibration

Calibration is a way of assessing probability statements against some idea of *truth*, a base measure $\mathbb{P}$.

- Truth is usually taken to be reality
- We are well-calibrated if $p\%$ of all predictions reported at probability $p$ are true. Idea goes back at least to Dawid (1984).
- Calibration is a difficult idea for Bayesians (subjectivists), even when $\mathbb{P}$ is reality's measure.
  - Seidenfeld (1985) wrote

> *'Calibration in the long run is otiose, and in the short run is an inducement to hedging'.*

Calibration, when $\mathbb{P}$ is reality, is a desirable frequency property to possess, but Bayesian's should beware of consciously aiming for it.

# Calibration

The definition of calibration used in FP is superficially similar:

- $\mathbb{P}_{ABC}$ is well calibrated if

$$\mathbb{P}(\theta \in A | E_q(A)) = q$$

where $E_q(A)$ is the event that the ABC posterior assigns probability $q$ to event $A$

i.e., given that $A$ is an event assigned probability $q$ by $\mathbb{P}_{ABC}$, then we are calibrated if $A$ occurs with probability $q$ according to base measure $\mathbb{P}$.

# Calibration

The definition of calibration used in FP is superficially similar:

- $\mathbb{P}_{ABC}$ is well calibrated if

$$\mathbb{P}(\theta \in A | E_q(A)) = q$$

where $E_q(A)$ is the event that the ABC posterior assigns probability $q$ to event $A$

i.e., given that $A$ is an event assigned probability $q$ by $\mathbb{P}_{ABC}$, then we are calibrated if $A$ occurs with probability $q$ according to base measure $\mathbb{P}$.

The difference with the standard definition, is the definition of the *base-measure*, $\mathbb{P}$. In FP's definition:

- $\mathbb{P}$ does not represent reality.
- It is defined by the prior, simulator, and summary used.
  - i.e., this definition of calibration ensures you are calibrated against your own beliefs.
  - The prior is calibrated under this definition.

Further, noisy-ABC is calibrated only if we repeated the analysis with multiple noisy datasets.