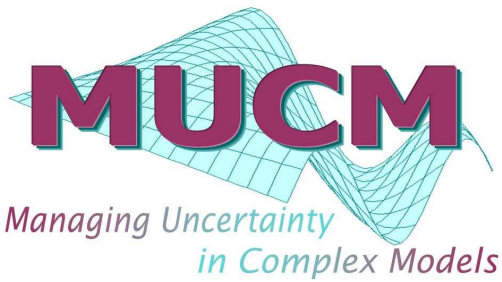
Análise Estatística de Simuladores
Lecture 6: Approximate Bayesian computation (ABC)

Leonardo Bastos[1] and Richard Wilkinson[2]

[1]Universidade Federal Fluminense
[2]University of Nottingham, UK

19º SINAPE, São Pedro, 27 July, 2010



The University of Nottingham

Mini-course lecture notes

All the slides from the course will be made available at
`www.professores.uff.br/lbastos`

or from

`www.maths.nottingham.ac.uk/personal/pmzrdw`

Please email us with questions, comments, suggestions for improvements etc!

Contents

This lecture is all about a Monte Carlo method of calibration known as ABC (approximate Bayesian computation).

1. Stochastic modelling
2. ABC algorithm
3. Approximate MCMC
4. Dating primate divergence times

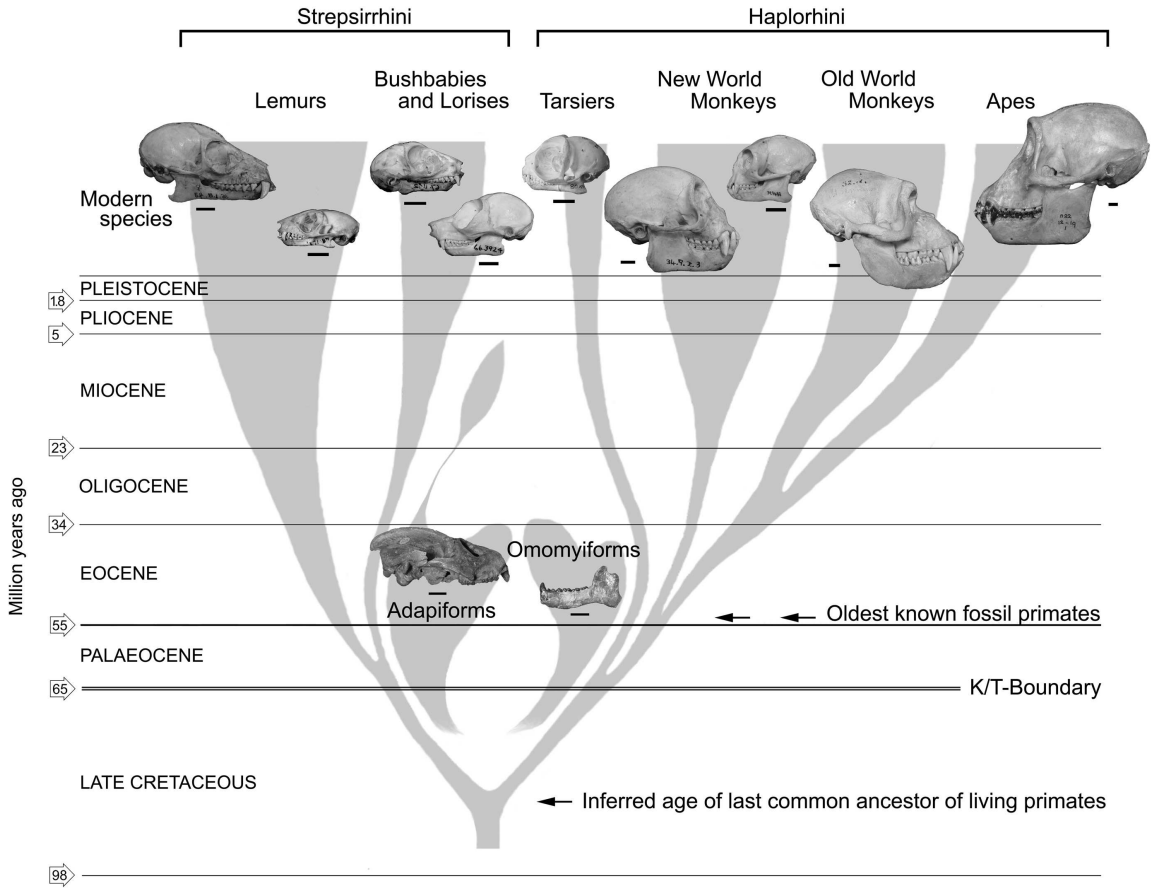
Stochastic modelling

Fitting to data

For forwards models we specify parameters θ and i.c.s and the model generates output X. Usually, we are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.

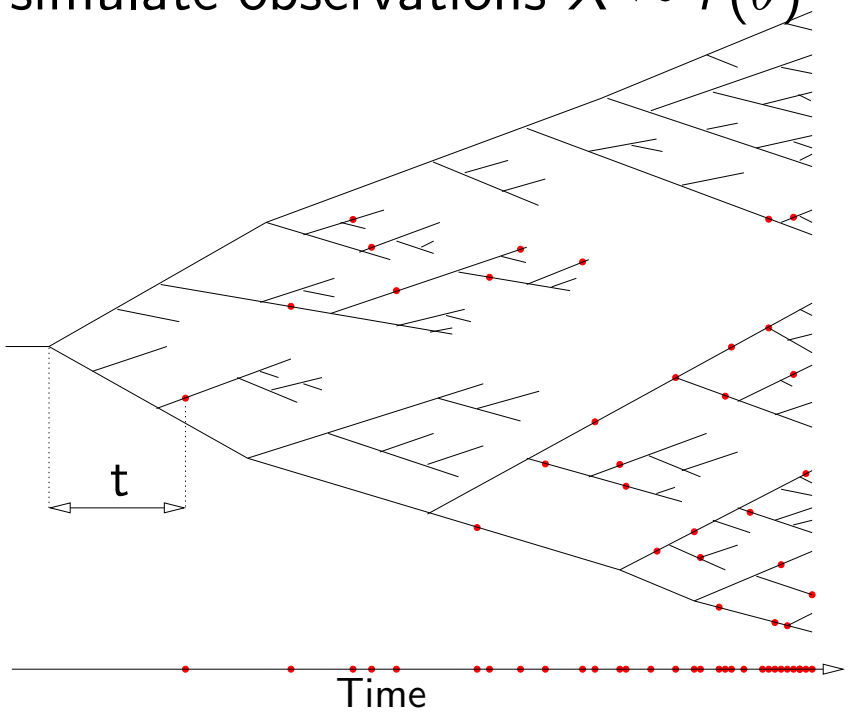
Different terminology:

- Calibration
- Data assimilation
- Parameter estimation
- Inverse-problem
- Bayesian inference



We now switch attention to stochastic computer models. Two types of statistical model:

- Prescribed models - likelihood function is specified $\pi(X|\theta)$
- Implicit models - mechanism to simulate observations $X \sim f(\theta)$

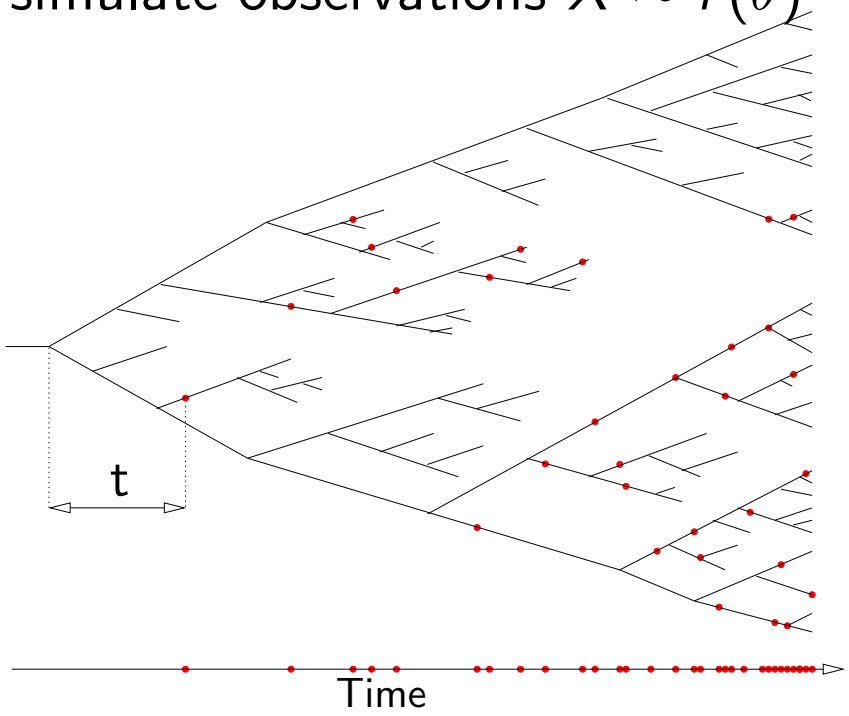


We now switch attention to stochastic computer models. Two types of statistical model:

- Prescribed models - likelihood function is specified $\pi(X|\theta)$
- Implicit models - mechanism to simulate observations $X \sim f(\theta)$

Implicit models give scientists more freedom to build complex models, and the increase in computer power has made their use more practicable.

Implicit models are popular in many disciplines, eg ecology, epidemiology, population biology, genetics.



Bayesian Inference

Aim to sample from the posterior distribution:

$$\pi(\theta|\mathcal{D}) \propto \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} = \frac{\pi(\theta)\pi(\mathcal{D}|\theta)}{\pi(\mathcal{D})}.$$

Bayesian Inference

Aim to sample from the posterior distribution:

$$\pi(\theta|\mathcal{D}) \propto \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} = \frac{\pi(\theta)\pi(\mathcal{D}|\theta)}{\pi(\mathcal{D})}.$$

- Monte Carlo methods enable Bayesian inference to be done in complex models

Bayesian Inference

Aim to sample from the posterior distribution:

$$\pi(\theta|\mathcal{D}) \propto \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} = \frac{\pi(\theta)\pi(\mathcal{D}|\theta)}{\pi(\mathcal{D})}.$$

- Monte Carlo methods enable Bayesian inference to be done in complex models
- The basic idea behind Monte Carlo methods is that instead of analytically calculating quantities of interest using the density $\pi(\theta)$, we instead sample from it and use the samples to approximate.

Bayesian Inference

Aim to sample from the posterior distribution:

$$\pi(\theta|\mathcal{D}) \propto \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} = \frac{\pi(\theta)\pi(\mathcal{D}|\theta)}{\pi(\mathcal{D})}.$$

- Monte Carlo methods enable Bayesian inference to be done in complex models
- The basic idea behind Monte Carlo methods is that instead of analytically calculating quantities of interest using the density $\pi(\theta)$, we instead sample from it and use the samples to approximate.
- For example, suppose we wish to find

$$\mathbb{E}f(\Theta) = \int f(\theta)\pi(\theta)\mathrm{d}\theta$$

If we can generate iid samples $\theta_1, \ldots, \theta_N \sim \pi(\theta)$ then we can use the Monte Carlo approximation

$$\mathbb{E}f(\Theta) \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)$$

- Standard MCMC can be difficult or impossible in many stochastic models, e.g., if

- Standard MCMC can be difficult or impossible in many stochastic models, e.g., if
  - $\pi(\mathcal{D}|\theta) = \frac{q(\mathcal{D}|\theta)}{C(\theta)}$ where $C(\theta)$ is unknown - doubly intractable distribution, e.g., Ising model.

- Standard MCMC can be difficult or impossible in many stochastic models, e.g., if
  - $\pi(\mathcal{D}|\theta) = \frac{q(\mathcal{D}|\theta)}{C(\theta)}$ where $C(\theta)$ is unknown - doubly intractable distribution, e.g., Ising model.
  - $\pi(\mathcal{D}|\theta)$ not even known up to a normalizing constant - completely intractable distribution.

- Standard MCMC can be difficult or impossible in many stochastic models, e.g., if
  - $\pi(\mathcal{D}|\theta) = \frac{q(\mathcal{D}|\theta)}{C(\theta)}$ where $C(\theta)$ is unknown - doubly intractable distribution, e.g., Ising model.
  - $\pi(\mathcal{D}|\theta)$ not even known up to a normalizing constant - completely intractable distribution.
  - completely intractable distributions can occur when you have highly dependent data arising from an underlying tree or graphical structure.
    - Population Genetics
    - Epidemiology
    - Evolutionary Biology

Approximate Bayesian Computation

Approximate Bayesian Computation

The challenged faced in computer experiments is that we do not typically know the likelihood function, and must instead do inference using only simulation from the model.

Approximate Bayesian Computation

The challenged faced in computer experiments is that we do not typically know the likelihood function, and must instead do inference using only simulation from the model.

Approximate Bayesian computation (ABC) algorithms are a group of Monte Carlo algorithms used for finding posterior distributions

- they do not require explicit knowledge of the likelihood function
- inference is done with draws from the model
- sometimes called 'likelihood-free' inference.

Approximate Bayesian Computation

The challenged faced in computer experiments is that we do not typically know the likelihood function, and must instead do inference using only simulation from the model.

Approximate Bayesian computation (ABC) algorithms are a group of Monte Carlo algorithms used for finding posterior distributions

- they do not require explicit knowledge of the likelihood function
- inference is done with draws from the model
- sometimes called 'likelihood-free' inference.

ABC methods are becoming popular in the biological sciences.

- Genetics - Siegmund *et al.* (DNA methylation in cancer cells); Foll *et al.* (using polymorphism markers to infer pop structure)
- Epidemiology - Blum and Tran (Cuban HIV-AIDS); Tanaka *et al.* (Tuberculosis)
- Population biology - Ratmann *et al.* (protein networks); Cornuet *et al.* (population history).

Likelihood-Free Inference

Rejection Algorithm

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

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

Likelihood-Free Inference

Rejection Algorithm

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

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

'Mechanical' Rejection Algorithm

- Draw θ from $\pi(\cdot)$
- Simulate $X \sim f(\theta)$
- Accept θ if $\mathcal{D} = X$

Theorem: The rejection algorithm samples from the posterior

Proof.

Theorem: The rejection algorithm samples from the posterior

Proof.

$$\text{Let } I = \begin{cases} 1 & \mathcal{D} = X \\ 0 & \mathcal{D} \neq X \end{cases}$$

Theorem: The rejection algorithm samples from the posterior

Proof.

$$\text{Let } I = \begin{cases} 1 & \mathcal{D} = X \\ 0 & \mathcal{D} \neq X \end{cases}$$

$$\text{Then} \qquad \mathbb{P}(I = 1|\theta) = \pi(\mathcal{D}|\theta)$$

$$\text{So} \qquad \mathbb{P}(I = 1) = \int \mathbb{P}(I = 1|\theta)\pi(\theta)\mathrm{d}\theta$$

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

Theorem: The rejection algorithm samples from the posterior

Proof.

$$\text{Let } I = \begin{cases} 1 & \mathcal{D} = X \\ 0 & \mathcal{D} \neq X \end{cases}$$

$$\text{Then} \qquad \mathbb{P}(I = 1|\theta) = \pi(\mathcal{D}|\theta)$$

$$\text{So} \qquad \mathbb{P}(I = 1) = \int \mathbb{P}(I = 1|\theta)\pi(\theta)\mathrm{d}\theta$$

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

$$\text{Hence} \qquad \pi(\theta|I = 1) = \frac{\mathbb{P}(I = 1|\theta)\pi(\theta)}{\mathbb{P}(I = 1)}$$

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

Acceptance rate

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

This opens the opportunity for calculating Bayes factors comparing two different models, η_1 and η_2.

$$B_{12} = \frac{\mathbb{P}_{\eta 1}(\mathcal{D})}{\mathbb{P}_{\eta 2}(\mathcal{D})}$$

Acceptance rate

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

This opens the opportunity for calculating Bayes factors comparing two different models, η_1 and η_2.

$$B_{12} = \frac{\mathbb{P}_{\eta 1}(\mathcal{D})}{\mathbb{P}_{\eta 2}(\mathcal{D})}$$

Estimate $\mathbb{P}_{\eta 1}(\mathcal{D})$ by the acceptance rate when using η_1 etc. See Xavier Didelot's work for more details.

Approximate Bayesian Computation I

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

Approximate Rejection Algorithm

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

Approximate Bayesian Computation I

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

Approximate Rejection Algorithm

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

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

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

δ reflects the tension between computability and accuracy.

Approximate Bayesian Computation II

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

Approximate Rejection Algorithm With Summaries

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

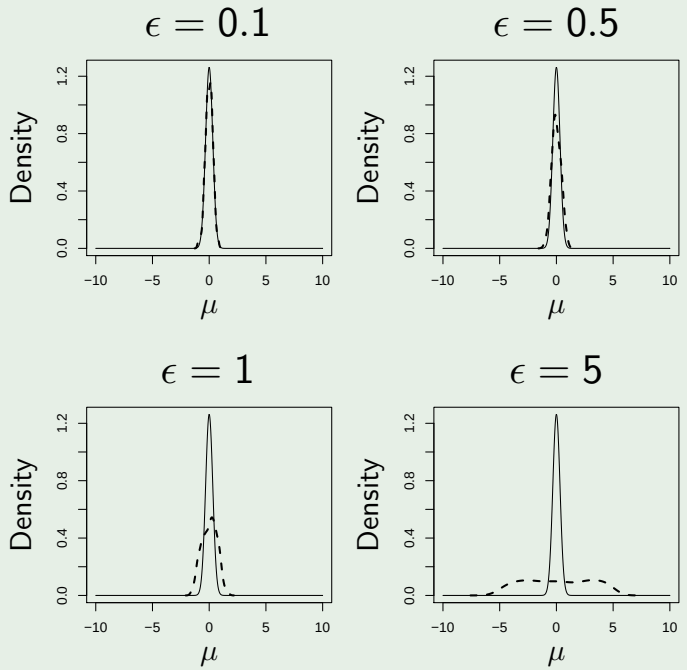
If S is sufficient this is equivalent to the previous algorithm (see Paul Fearnhead and Dennis Prangle's work for more details).

There are many other extensions: approximate MCMC, approximate SMC, approximate particle-MCMC, etc.

Example (Gaussian Distribution)

Suppose $X_i \sim N(\mu, \sigma^2)$, with σ^2 known, and give μ an improper flat prior distribution, $\pi(\mu) = 1$ for $\mu \in \mathbb{R}$.

1000 samples



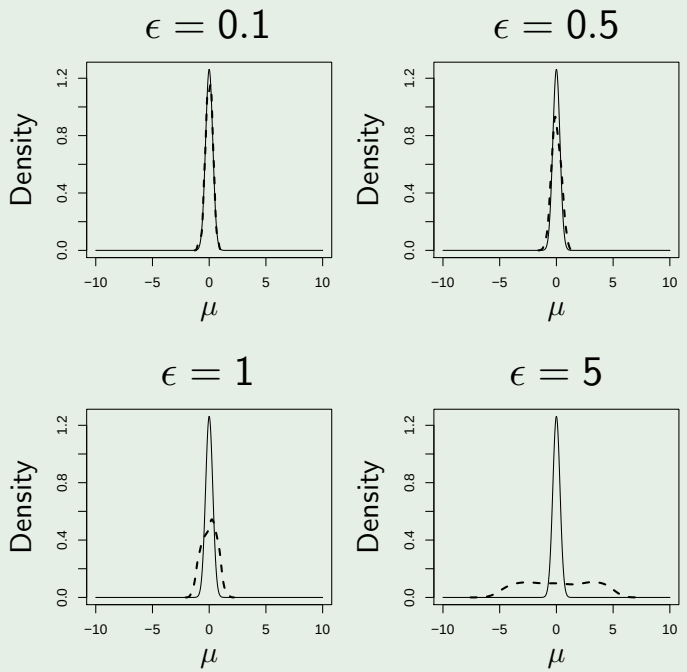
Example (Gaussian Distribution)

Suppose $X_i \sim N(\mu, \sigma^2)$, with σ^2 known, and give μ an improper flat prior distribution, $\pi(\mu) = 1$ for $\mu \in \mathbb{R}$.

Suppose we observe data with $\bar{x} = 0$.

- Pick $\mu \sim U(-\infty, \infty)$
- Simulate $X_i \sim N(\mu, \sigma^2)$
- Accept μ if $|\bar{x}| < \epsilon$.

1000 samples



Example (Gaussian Distribution)

Suppose $X_i \sim N(\mu, \sigma^2)$, with σ^2 known, and give μ an improper flat prior distribution, $\pi(\mu) = 1$ for $\mu \in \mathbb{R}$.

Suppose we observe data with $\bar{x} = 0$.

- Pick $\mu \sim U(-\infty, \infty)$
- Simulate $X_i \sim N(\mu, \sigma^2)$
- Accept μ if $|\bar{x}| < \epsilon$.

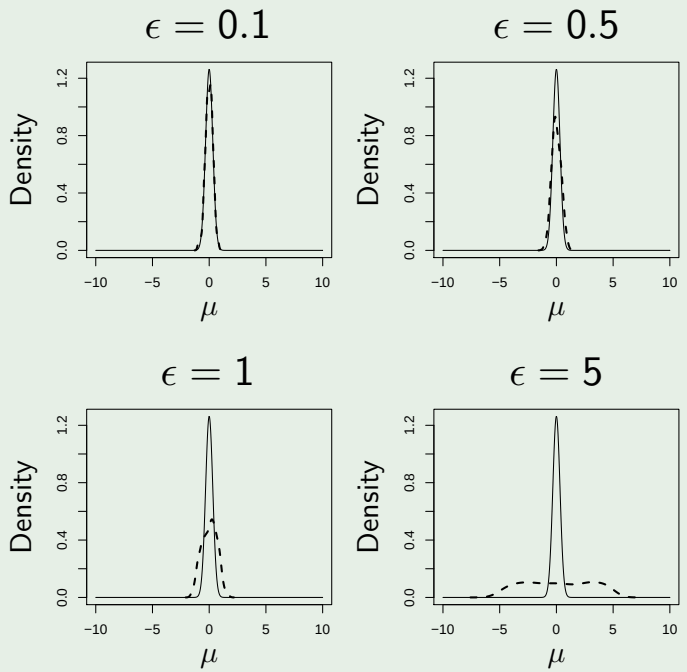
Then $\pi(\mu \mid |\bar{x}| \leq \epsilon) =$

$$\frac{\Phi\left(\frac{\epsilon-\mu}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\frac{-\epsilon-\mu}{\sqrt{\sigma^2/n}}\right)}{2\epsilon}$$

and

$$\mathbb{V}ar(\mu \mid |\bar{x}| \leq \epsilon) = \mathbb{V}ar(\mu \mid \bar{x} = 0) + \frac{\epsilon^2}{3}$$

1000 samples



Markov chain Monte Carlo (MCMC)

Metropolis-Hastings Algorithm

Lets briefly return to the world of likelihood based inference (prescribed models).

Metropolis-Hastings Algorithm

Lets briefly return to the world of likelihood based inference (prescribed models).

Rejection sampling is inefficient, as θ is repeatedly sampled from its prior distribution.

- The idea behind MCMC is that by correlating observations more time is spent in regions of high likelihood.

Metropolis-Hastings Algorithm

Lets briefly return to the world of likelihood based inference (prescribed models).

Rejection sampling is inefficient, as θ is repeatedly sampled from its prior distribution.

- The idea behind MCMC is that by correlating observations more time is spent in regions of high likelihood.

Suppose we wish to simulate from a (possibly multivariate) distribution $\pi(x)$. We need to find a transition kernel $Q(x, y)$ such that

- Simulation of a Markov chain $\{X_1, X_2, \ldots\}$ with transition kernel Q is straightforward.
- The stationary distribution of the Markov chain is $\pi(x)$.

Metropolis-Hastings Algorithm

Lets briefly return to the world of likelihood based inference (prescribed models).

Rejection sampling is inefficient, as θ is repeatedly sampled from its prior distribution.

- The idea behind MCMC is that by correlating observations more time is spent in regions of high likelihood.

Suppose we wish to simulate from a (possibly multivariate) distribution $\pi(x)$. We need to find a transition kernel $Q(x, y)$ such that

- Simulation of a Markov chain $\{X_1, X_2, \ldots\}$ with transition kernel Q is straightforward.
- The stationary distribution of the Markov chain is $\pi(x)$.

The Metropolis-Hastings algorithm tells us how to build transition kernels such that the Markov chain converges to the distribution $\pi(x)$.

Metropolis-Hastings Algorithm - II

Metropolis-Hastings Algorithm

1. Choose a starting location $X_0 = x_0$.
2. Suppose at time t, we have $X_t = x$. Propose a candidate value y from proposal distribution $q(x, y)$.
3. Calculate the 'Metropolis-Hastings acceptance probability' $\alpha(x, y)$ given by

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

4. Accept the move with probability $\alpha(x, y)$. If the move is rejected, stay at x. That is, we set

$$X_{t+1} = \begin{cases} y & \text{with probability } \alpha(x, y) \\ x & \text{with probability } 1 - \alpha(x, y) \end{cases}$$

Approximate MCMC for Bayesian inference

Approximate MCMC for Bayesian inference

Metropolis-Hastings(MCMC) Algorithm

- Suppose we are currently at θ. Propose θ' from density $q(\theta, \theta')$.
- Calculate

$$h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\pi(D|\theta')q(\theta', \theta)}{\pi(\theta)\pi(D|\theta)q(\theta, \theta')}\right).$$

- Accept the move to θ' with probability $h(\theta, \theta')$, else stay at θ.

This will simulate a Markov chain $\{\theta_1, \theta_2, \ldots\}$ which has stationary distribution $\pi(\theta|\mathcal{D})$.

Approximate MCMC (Marjoram *et al.*)

However, we are assuming that we have a complex computer model and don't know the likelihood function $\pi(\mathcal{D}|\theta)$.

Approximate MCMC (Marjoram *et al.*)

However, we are assuming that we have a complex computer model and don't know the likelihood function $\pi(\mathcal{D}|\theta)$.

Instead, we can approximate this ratio using simulation giving us the approximate MCMC algorithm.

Approximate Metropolis-Hastings Algorithm

- Suppose we are currently at θ. Propose θ' from density $q(\theta, \theta')$.
- Simulate X from $f(\theta')$.
- If $\rho(\mathcal{D}, X) \leq \epsilon$, calculate

$$h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right).$$

- Accept the move to θ' with probability $h(\theta, \theta')$, else stay at θ.

$\exists$ approximate sequential Monte Carlo methods (see Sisson *et al.*).

Why are ABC methods popular?

- Likelihood not needed
- Trivial to code - embarrassingly paralizable
- Less tuning required than in MCMC
- Very little user input required

Why are ABC methods popular?

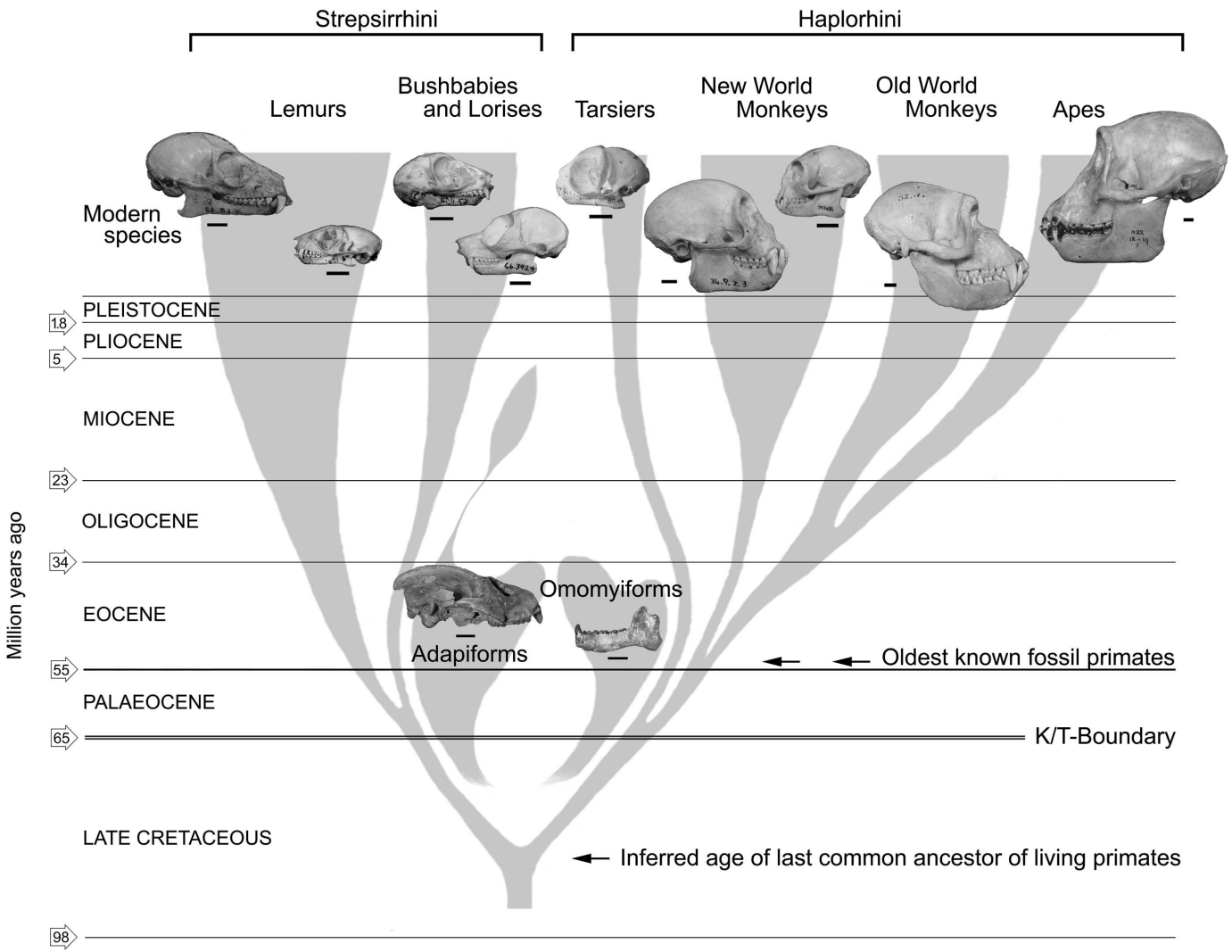
- Likelihood not needed
- Trivial to code - embarrassingly paralizable
- Less tuning required than in MCMC
- Very little user input required

However, despite numerous uses in application papers, there are many unresolved issues with ABC methods

- The methods are approximate with the quality of the approximation unknown in general
  - Wilkinson 2009 partially answers this question.
- Several implementation decisions are required. Choice of
  - metric
  - tolerance
  - summary statistic, etc

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:

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

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:

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.

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.

Who cares?

Why does this gap matter?

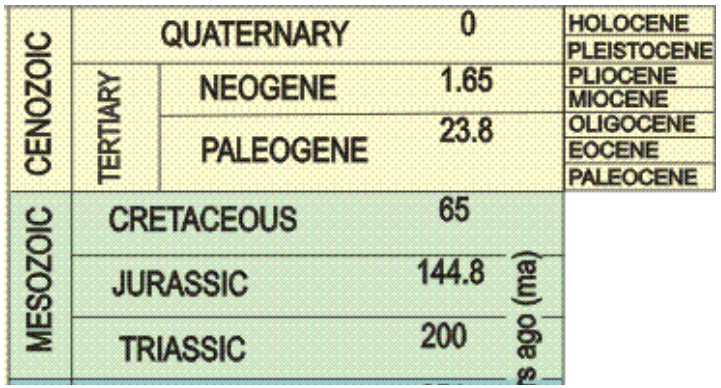
- Primates provide the zoological context for human evolution
  - has consequences for the human-chimp divergence time.

Who cares?

Why does this gap matter?

- Primates provide the zoological context for human evolution
  - has consequences for the human-chimp divergence time.
- Did primates coexist with the dinosaurs?

- Dinosaurs became extinct at the K-T crash (cretaceous-tertiary boundary) 65 mya.
- Did primates evolve to fill an evolutionary niche?

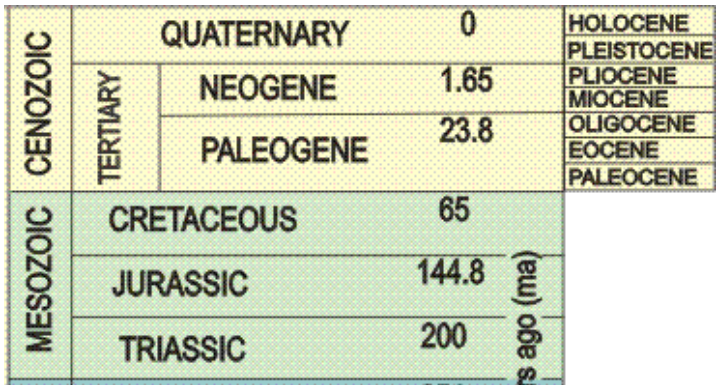


Who cares?

Why does this gap matter?

- Primates provide the zoological context for human evolution
  - has consequences for the human-chimp divergence time.
- Did primates coexist with the dinosaurs?

- Dinosaurs became extinct at the K-T crash (cretaceous-tertiary boundary) 65 mya.
- Did primates evolve to fill an evolutionary niche?

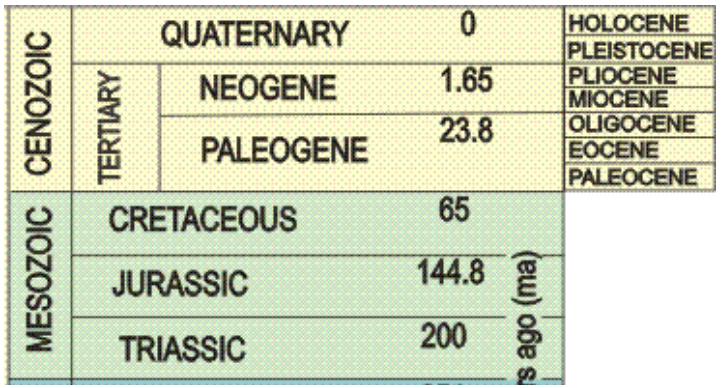


- Timing of primate evolution relative to evolution of angiosperms and large fruit bearing plants.

Who cares?

Why does this gap matter?

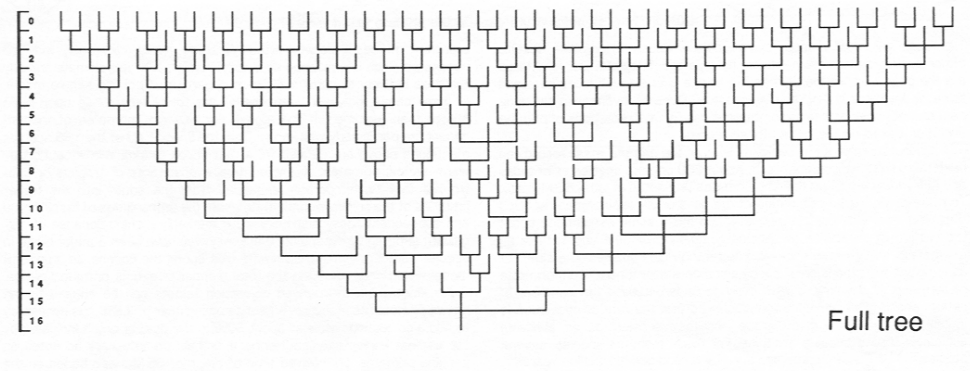
- Primates provide the zoological context for human evolution
  - has consequences for the human-chimp divergence time.
- Did primates coexist with the dinosaurs?



- Dinosaurs became extinct at the K-T crash (cretaceous-tertiary boundary) 65 mya.
- Did primates evolve to fill an evolutionary niche?
  - Timing of primate evolution relative to evolution of angiosperms and large fruit bearing plants.
  - Divergence time has consequences for theories about the dispersion of primates to the new world.

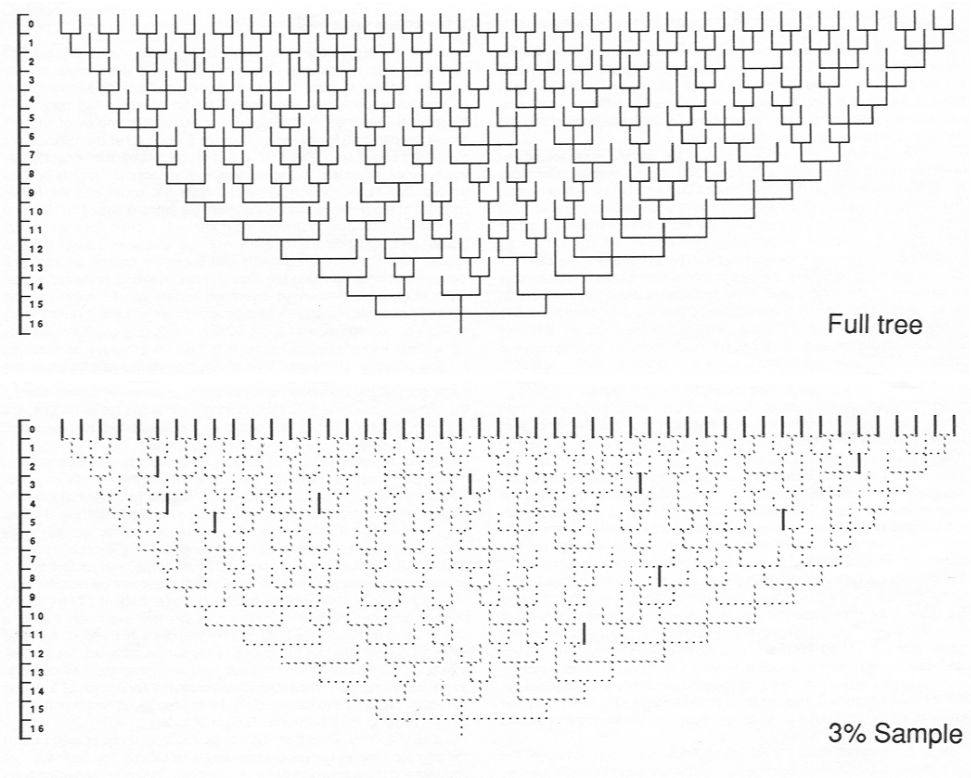
Why is this difficult?

Non-repeatable event



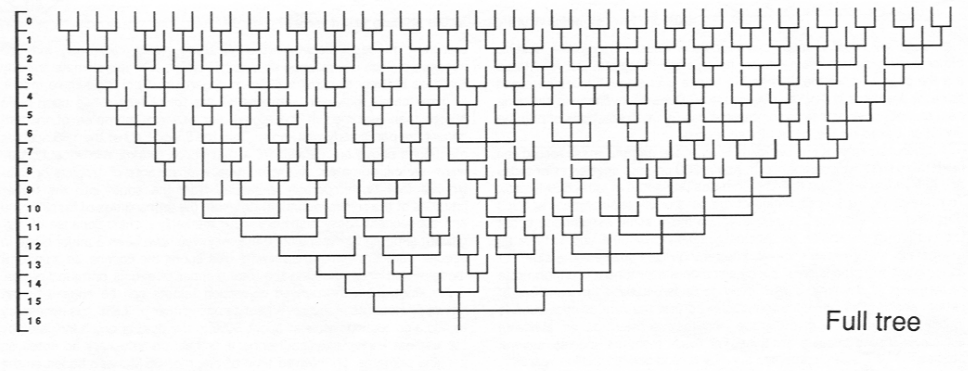
Why is this difficult?

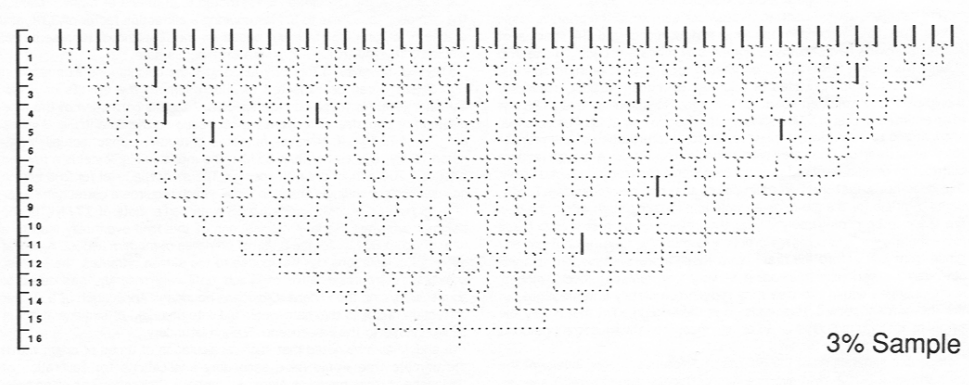
Non-repeatable event

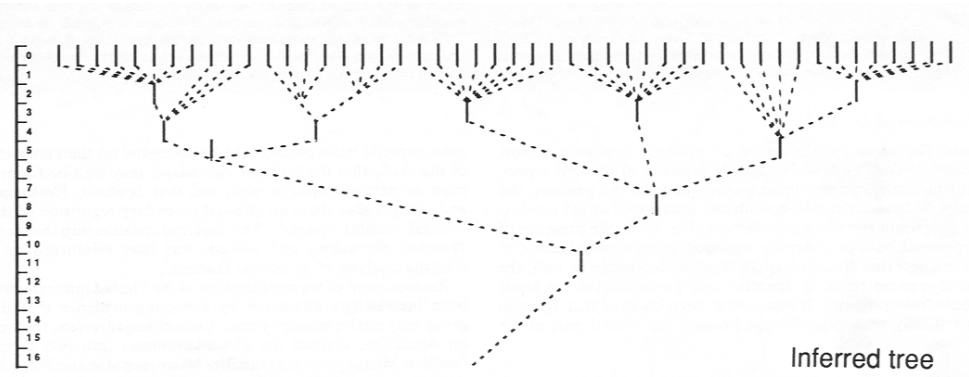


Why is this difficult?

Non-repeatable event







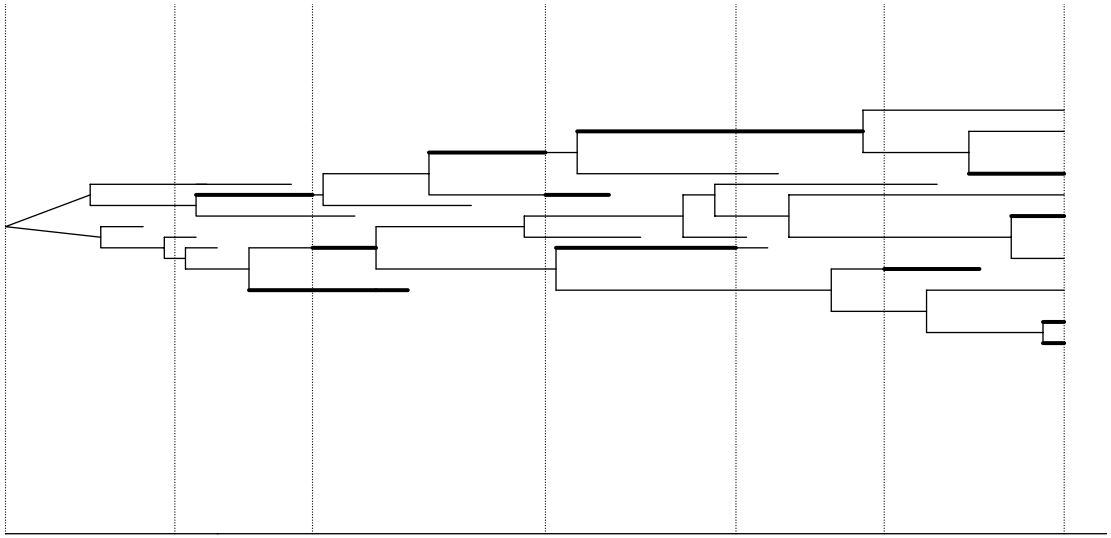
Data

Robert Martin and Christophe Soligo

Epoch	k	Time at base of Interval k	Primate fossil counts (D_k)	Anthropoid fossil counts (S_k)
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



Used an inhomogeneous binary Markov branching process 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 $m(u) = \mathbb{E}L_u$.

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

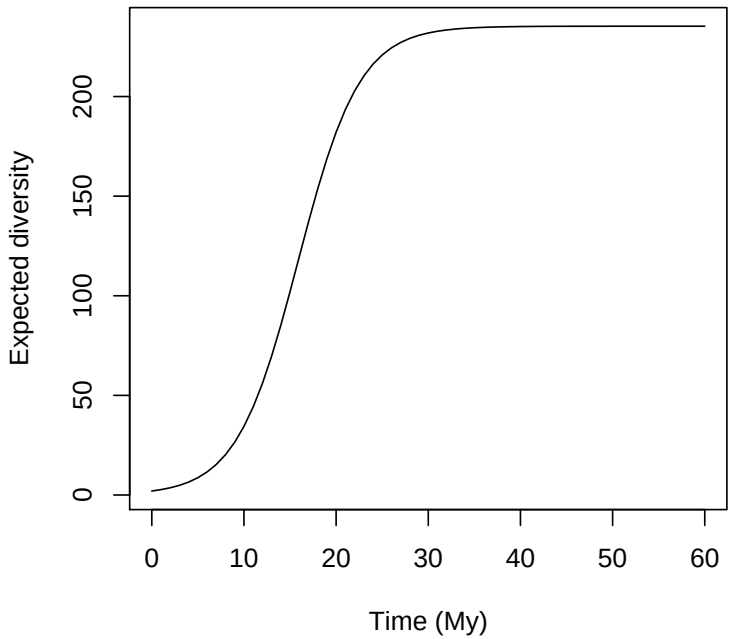
$$\mathbb{E}(Z(t) = n | Z(0) = 2) = 2 \exp\left(\lambda \int_0^t (m(u) - 1) du\right)$$

  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

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

- Poisson Model: fossils are superimposed on all branches contained in epoch i as a Poisson process of rate β_i.

$$\mathbb{P}(D_i|\mathcal{T}) \sim Po(\beta_i l_i)$$

  where l_i is the total length of all lineages in epoch i.

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

- Poisson Model: fossils are superimposed on all branches contained in epoch i as a Poisson process of rate β_i.

$$\mathbb{P}(D_i|\mathcal{T}) \sim Po(\beta_i l_i)$$

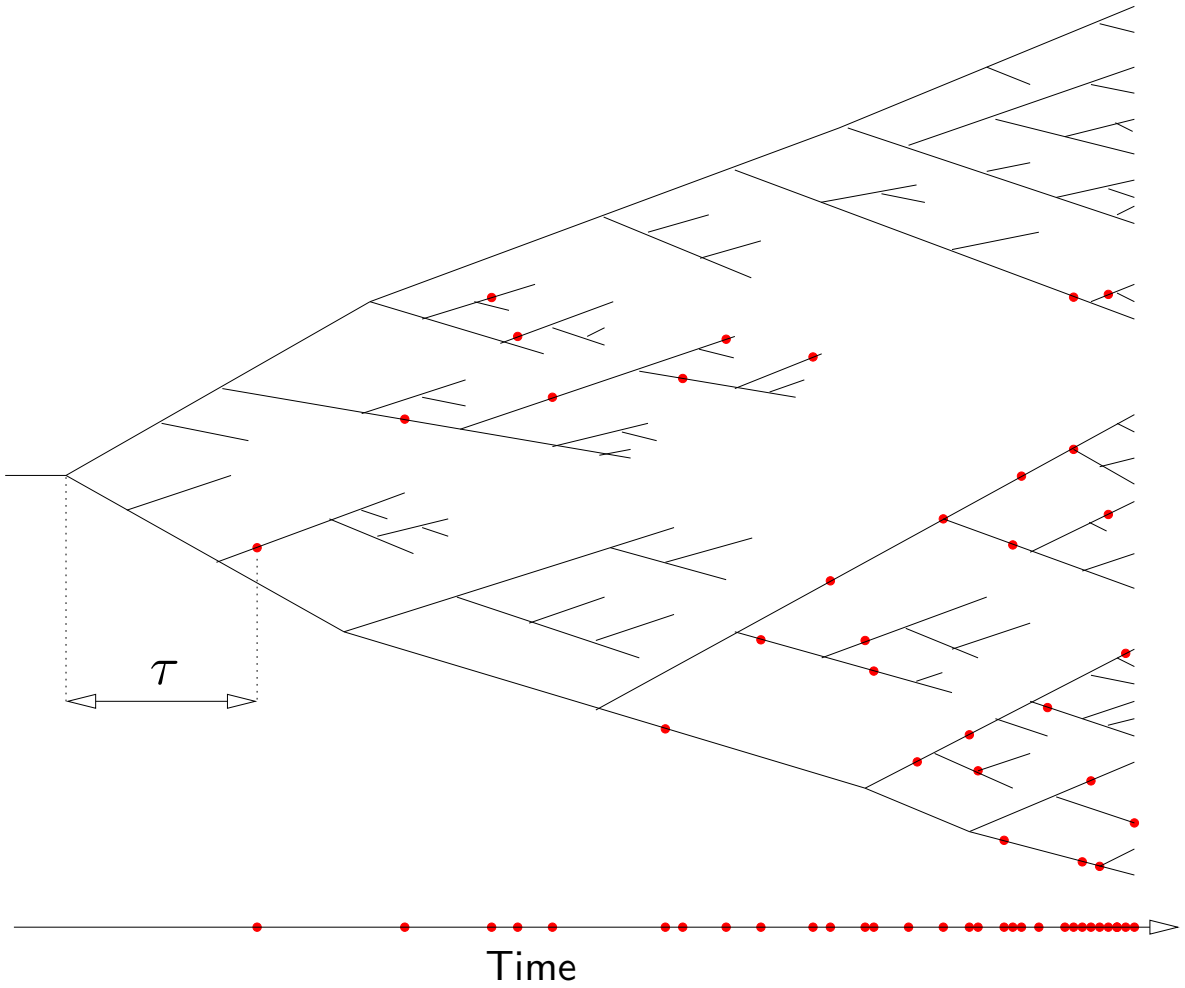
  where l_i is the total length of all lineages in epoch i.

NOTE: Sampling rates vary for each epoch - accounts for variable sampling, e.g., pull of the recent

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.

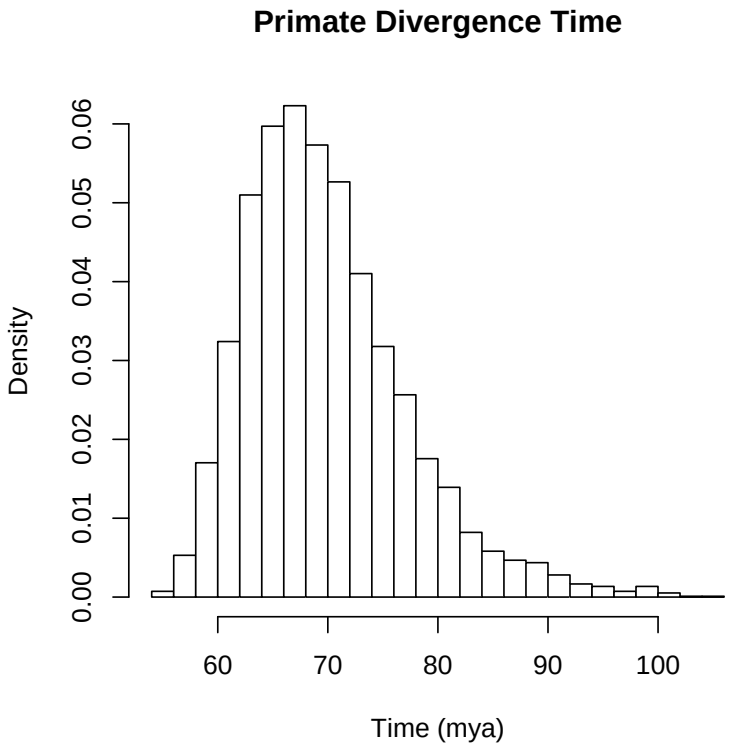
ABC choices

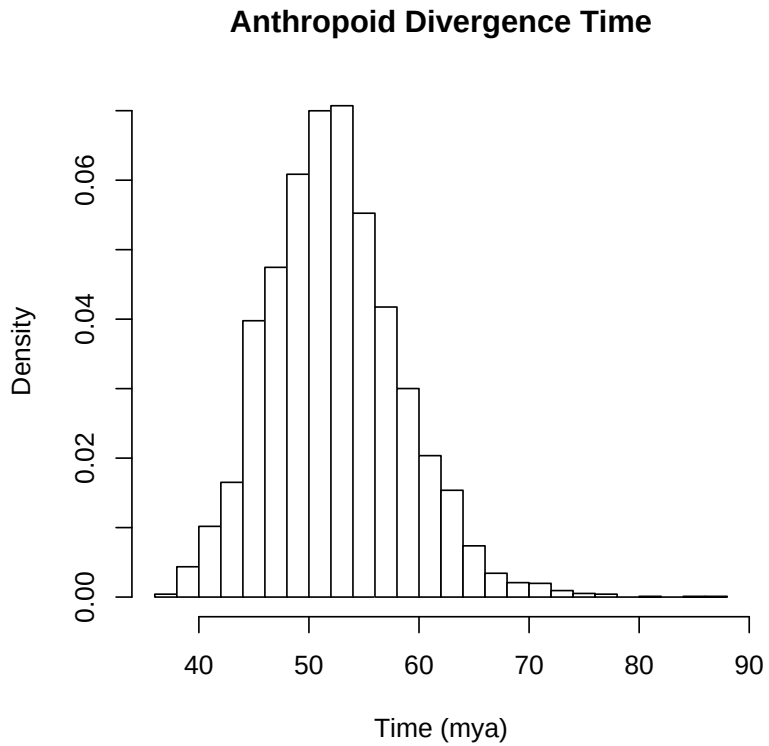
- Metric

$$\rho(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{D'_i}{D'_+} \right| + \left| \frac{D'_+}{D_+} - 1 \right| + \left| \frac{N'_0}{N_0} - 1 \right|$$

- Tolerance: use the smallest feasible value $\delta = 0.1$
- No summaries here

Results





Modelling the K-T crash causes shorter tails in the posterior distribution of the divergence times.

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 florensis (the hobbit man), or a microcephalic human, etc?
- 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.

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
- Etc.

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 diversity, ignoring these smaller fluctuations which are likely to be important for the preservation of fossils in well sampled regions.

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.
- Similarly, large D_i presumably have a more variable 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.

Choice of metric

Our analysis used a metric which contained the total variation distance between the two sets of proportions

$$\rho(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{D'_i}{D'_+} \right| + \left| \frac{D'_+}{D_+} - 1 \right| + \left| \frac{N'_0}{N_0} - 1 \right|$$

We can rewrite the variation distance part as

$$\sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{D'_i}{D'_+} \right| = \sum \left| \frac{D'_i}{D'_+} \right| \left| \frac{D'_+}{D_+} \frac{D_i}{D'_i} - 1 \right|$$

so that the contribution from each epoch is weighted by $\left| D'_i / D'_+ \right|$.

Unfortunately, we can see that $\frac{D'_+}{D_+}$ occurs in every weight, which is undesirable.

Also, each term is given equal weight, hence assuming an equal level of accuracy.

An improved metric

In theory, we can account for all these issues by using the generalised ABC algorithm. We could use an acceptance probability of the form

$$\pi_\epsilon(\mathcal{D}') = \pi(N_0'|N_0) \prod_{i=1}^{14} \pi_i(D_i'|D_i)$$

where $\pi_i(D_i'|D_i)$ is the density of the error on D_i.
In practice, this is a difficult exercise in elicitation in order to specify the errors, and to convolve all the different sources of error.

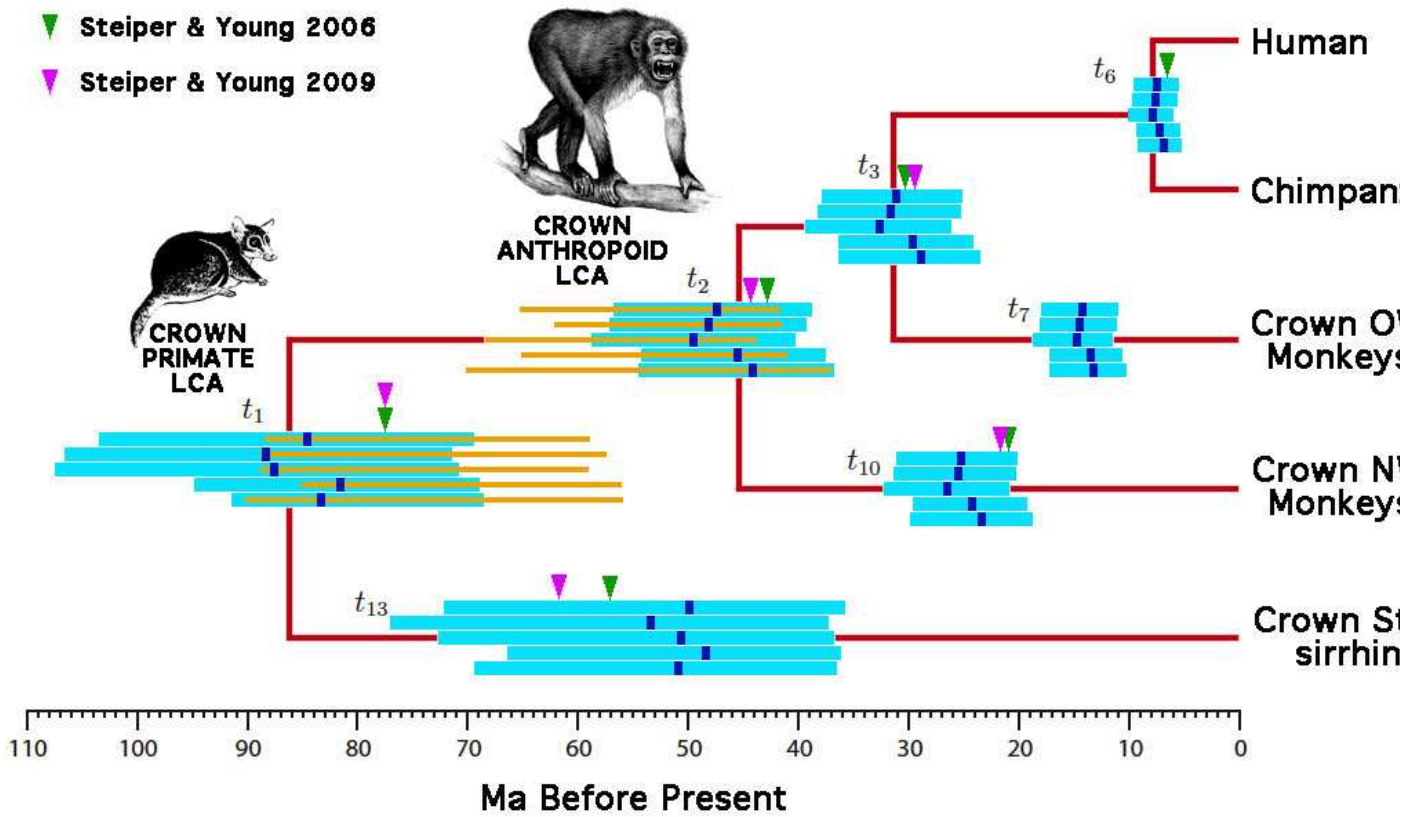
It is also a difficult computational challenge. Two ideas that help:

- We can use the fact that we know the distribution of D_i given N_i, to help break down the problem (e.g. we could accept the tree with probability $\prod \pi(D_i|N_i)$).
- Sequential particle methods can be used (plain SMC, not ABC-SMC).

Results of joint analysis

Wilkinson, Steiper, etal 2010

We can combine fossil divergence time estimates with genetic evidence to estimate the shape of the primate phylogeny



Extension results

Using just the fossil record, we find

- 95% CI for primate divergence time is $[83.4, 54.8]$ Mya.
- 95% CI for anthropoid divergence time is $[52, 37]$ Mya.

Extension results

Using just the fossil record, we find

- 95% CI for primate divergence time is $[83.4, 54.8]$ Mya.
- 95% CI for anthropoid divergence time is $[52, 37]$ Mya.

Combining these estimate with dna evidence we find

- 95% CI for primate divergence time is $[106.8, 69.2]$ Mya
- 95% CI for anthropoid divergence time is $[56.4, 38.8]$ Mya.
- the human-chimp divergence occured $[9.6, 5.7]$ Mya.

Summary

Approximate Bayesian computation (ABC) algorithms provide a likelihood-free approach to inference.

- ABC methods directly approximate the posterior distribution, whereas we use emulators to approximate the simulator.
- ABC methods are trivial to apply and can be used on any cheap stochastic model
- Care must be taken to ensure that a sensible tolerance and metric are used.

Summary

Approximate Bayesian computation (ABC) algorithms provide a likelihood-free approach to inference.

- ABC methods directly approximate the posterior distribution, whereas we use emulators to approximate the simulator.
- ABC methods are trivial to apply and can be used on any cheap stochastic model
- Care must be taken to ensure that a sensible tolerance and metric are used.

We hope that this course has given you a quick taste of some of the challenges and methods that are being studied in the field of computer experiments.

It is a new area of statistics, but is growing rapidly. Working in the field gives you an opportunity to collaborate with scientists from a range of disciplines.

Summary

Approximate Bayesian computation (ABC) algorithms provide a likelihood-free approach to inference.

- ABC methods directly approximate the posterior distribution, whereas we use emulators to approximate the simulator.
- ABC methods are trivial to apply and can be used on any cheap stochastic model
- Care must be taken to ensure that a sensible tolerance and metric are used.

We hope that this course has given you a quick taste of some of the challenges and methods that are being studied in the field of computer experiments.

It is a new area of statistics, but is growing rapidly. Working in the field gives you an opportunity to collaborate with scientists from a range of disciplines.

Thank you for listening!

References

- M. A. Beaumont and W. Zhang and D. J. Balding (2002) *Approximate Bayesian Computatation in Population Genetics*. Genetics 162, 2025-2035.
- P. Marjoram and J. Molitor and V. Plagnol and S. Tavaré (2003) *Markov Chain Monte Carlo without likelihoods*. PNAS 100, 15324-15328.
- S. A. Sisson and Y. Fan and M. M. Tanaka (2007) *Sequential Monte Carlo without Likelihoods*. PNAS 104, 1760-1765.
- Wilkinson (2009) *Approximate Bayesian computation (ABC) gives exact results under the assumption of model error*. In submission. Available as 1. arXiv:0811.3355, 2009.
- Wilkinson and S. Tavaré (2009) *Estimating the primate divergence time using conditioned birth-and-death processes*, Theoretical Population Biology 75, pp. 278-285.
- Wilkinson, M. Steiper, C. Soligo, R.D. Martin, Z. Yang, and S. Tavaré (2010) *Dating primate divergences through an integrated analysis of palaeontological and molecular data*. In press, Systematic Biology.
- MUCM toolkit, available at `http://mucm.aston.ac.uk/MUCM/MUCMToolkit/`