Bayesian Statistics in Molecular Evolution

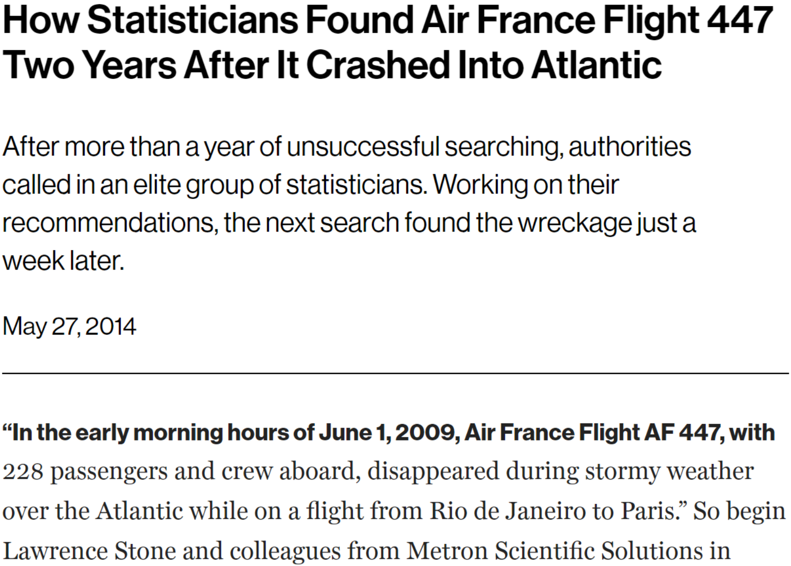
Yujin Chung

The Department of Mathematics & Statistics
Auburn University

Ewha Womans University

December 19, 2017

Missing Flight Found Using Bayes' Theorem



How Statisticians Found Air France Flight 447 Two Years After It Crashed Into Atlantic

After more than a year of unsuccessful searching, authorities called in an elite group of statisticians. Working on their recommendations, the next search found the wreckage just a week later.

May 27, 2014

"In the early morning hours of June 1, 2009, Air France Flight AF 447, with 228 passengers and crew aboard, disappeared during stormy weather over the Atlantic while on a flight from Rio de Janeiro to Paris." So begin Lawrence Stone and colleagues from Metron Scientific Solutions in

Missing Flight Found Using Bayes' Theorem

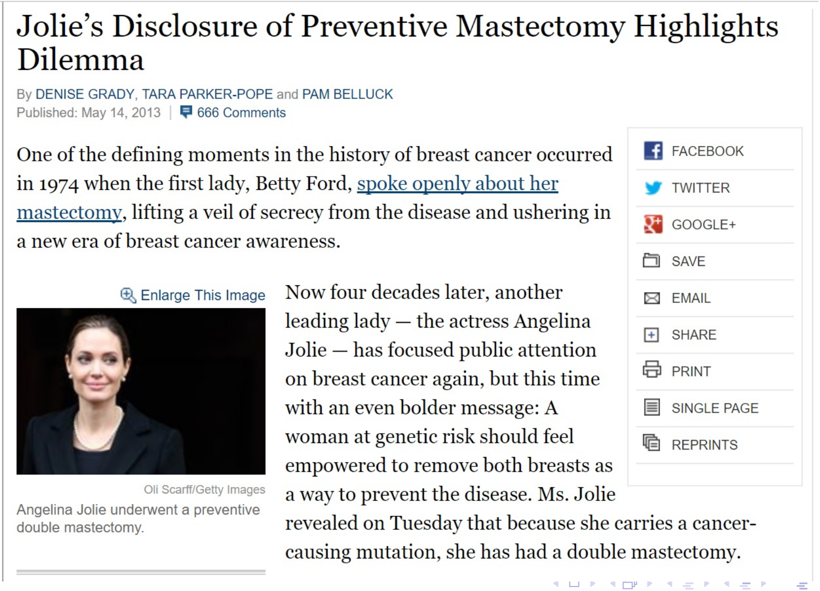
Data: the crash location, weather, flight record, searching record etc

Given the data, the probability of finding the wreckage at a given location

P(finding the wreckage at a given location | Data)

Used Bayesian inference to update the probability of finding the wreckage at a given location

Making a decision



Jolie's Disclosure of Preventive Mastectomy Highlights Dilemma

By DENISE GRADY, TARA PARKER-POPE and PAM BELLUCK
Published: May 14, 2013 | 666 Comments

One of the defining moments in the history of breast cancer occurred in 1974 when the first lady, Betty Ford, spoke openly about her mastectomy, lifting a veil of secrecy from the disease and ushering in a new era of breast cancer awareness.

Enlarge This Image

Oli Scarff/Getty Images

Angelina Jolie underwent a preventive double mastectomy.

Now four decades later, another leading lady — the actress Angelina Jolie — has focused public attention on breast cancer again, but this time with an even bolder message: A woman at genetic risk should feel empowered to remove both breasts as a way to prevent the disease. Ms. Jolie revealed on Tuesday that because she carries a cancer-causing mutation, she has had a double mastectomy.

Making a decision

The probability of having breast cancer: $\text{P(Having breast cancer)} = 12\%$

The probability of having breast cancer given personal information

- Data: Family history, BRCA1 mutant

$$\text{Pr(having breast cancer}|\text{Data)} = 87\%$$

The probability of having breast cancer after taking preventive mastectomy:

$$\text{Pr(having breast cancer}|\text{Preventive surgery)} = 5\%$$

Bayesian inference was used to update the probability of having breast cancer. Our decisions can be changed when the new data/information is provided.

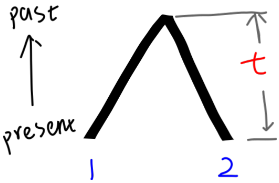
Bayesian inference in Molecular Evolution

Wildely applied to estimate

- phylogenetic trees (gene tree, species tree)
- sequence alignment
- species' divergence time
- demographic information (population size, migration rates etc)
- recombination breakpoints
- and more

Statistical inference

- θ: parameter of interest
  - ex) phylogenetic tree



- D: observed data
  - ex) DNA alignments

1 A A T T C C
2 C A T T G C

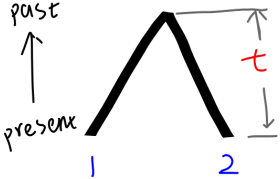
- Statistical inference:
  Estimating t (tree height) from the observed data set D

Traditional statistical inference

The probability of data set

$$\Pr(D|t)$$

- data are a repeatable random sample
- Parameters are constant (unknown)



1 A A T T C C

2 C A T T G C

Ex) Jukes-Cantor (JC) substitution model

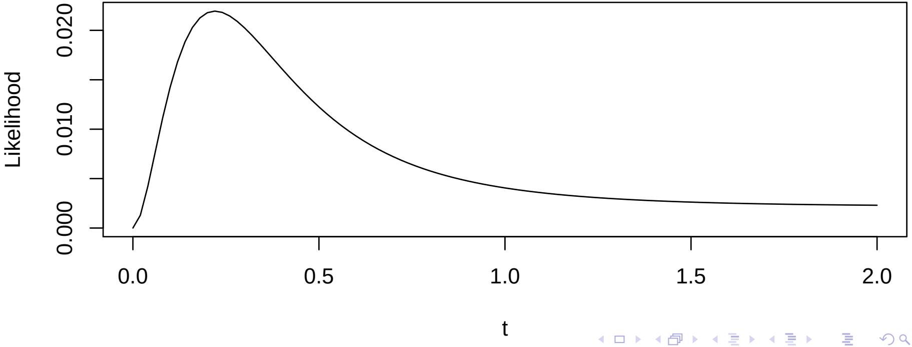
$$\Pr(D|t) \propto \left(\frac{3}{4} - \frac{3}{4}e^{-8t/3}\right)^2 \left(\frac{1}{4} + \frac{3}{4}e^{-8t/3}\right)^4$$

Likelihood

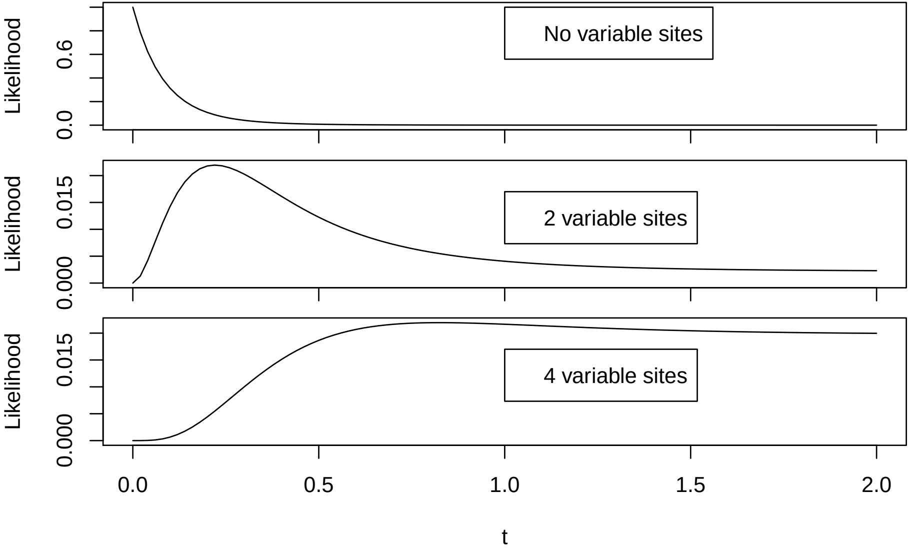
Likelihood:

$$L(t) = \left(\frac{3}{4} - \frac{3}{4}e^{-8t/3}\right)^2 \left(\frac{1}{4} + \frac{3}{4}e^{-8t/3}\right)^4$$

- function of parameter



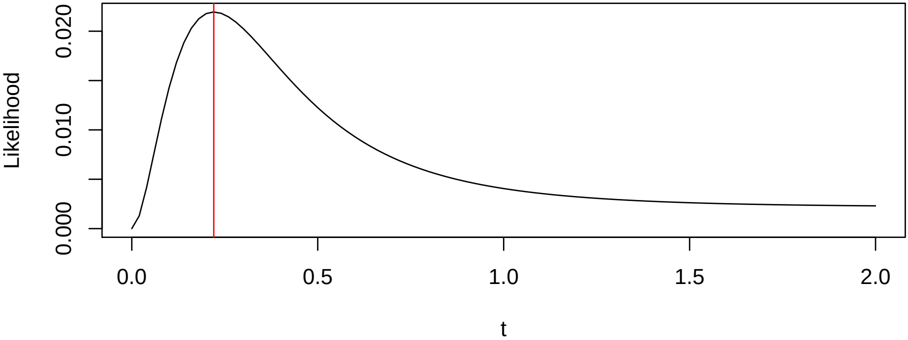
Likelihood gives some weight to all possible reconstructions



More variable sites, more weight on longer branch.

Maximum likelihood estimaton (MLE)

$$\hat{t} = \arg\max L(t)$$



The likelihood has the maximum value at $\hat{t} = 0.22042$.

Challenges in ML Tree Estimation

Many software: PAUP, PhyML, MEGA, . . .

- Parameters: tree topology and branch lengths
- Local maximum problem: no analytical solution for MLEs.
- Rapidly growing the space of parameter with the number of taxa

No. seq.	2	3	4	5	6	12	22
No. tree topo.	1	3	15	105	945	13 billion	1.3×10^{25}

Bayesian Inference

Goal: Estimating the posterior distribution

$$\underbrace{p(t \mid D)}_{\text{posterior distribution}} \propto \underbrace{p(D \mid t)}_{\text{Likelihood } L(t)} \times \underbrace{p(t)}_{\text{prior distribution}}$$

- **Prior distribution**: the distribution of parameter (e.g., branch length, t) before the data is provided
- **Posterior distribution**: the (updated) distribution of parameter (e.g., branch length, t) after the data is analyzed

- Example: computing the posterior distribution of tree (parameter of interest) using the observed DNA sequences (Data) and the prior distribution (prior knowlege)

Bayes' Theorem

Let A and B be events and $P(B) \neq 0$.

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)}. \end{aligned}$$

Interpretation

Bayes' Theorem:

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)}.$$

Let A be a parameter or hypothesis which we can't observe and B be the data or evidence which we can observe.

- $P(A)$: **prior probability** of A or initial belief in A. Event B is not accounted.
- $P(A|B)$: **posterior probability** of A given data B

Bayes' Theorem computes the posterior probility of A after observing B.

Interpretation

The crashing flight

- A: the wreckage location
- B: Data (flight record, weather etc)

Posterior probability of the wreckage location: $P(A|B)$

Breast cancer

- A: Breast cancer
- B: Data (family history, BRCA1)

Posterior probability of having breast cancer: $P(A|B)$

Example: What is the probability of cancer?

Approximately 1% of women aged 40-50 have breast cancer. Let's say women take a screening test (e.g., mammogram, genetic test). A woman with breast cancer has a 90% chance of a positive result, while a woman without has a 10% chance of a false positive result.

What is the probability a woman has breast cancer given that she just had a positive test?

Example: What is the probability of cancer?

Let A be the event that the woman has breast cancer and B that a positive result. We wish to calculate $P(A|B)$.

What we know from the problem:

- *"1% of women aged 40-50 have breast cancer"*
  Prior probability: $\mathrm{P(A)} = 0.01$
- *"A woman with breast cancer has a 90% chance of a positive result"*
  Conditional probability: $P(B|A) = 0.9$
- *"a woman without breast cancer has a 10% chance of a false positive result."*
  Conditional probability: $P(B|A^C) = 0.1$

Example: What is the probability of cancer?

Let A be the event that the woman has breast cancer and B that a positive result. What we know from the problem:
P(A) $= 0.01$, $P(B|A) = 0.9$ and $P(B|A^C) = 0.1$

We apply Bayes' theorem to compute $P(A|B)$:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} \\ &= 0.083 \end{aligned}$$

The prior probability of having breast cancer is 1%. Given a positive result of the screening test, however, **the posterior probability of breast cancer is 8.3%**.

Bayesian Inference

From Bayes' theorem, the posterior distribution is

$$p(t|D) = \frac{p(D|t)p(t)}{p(D)},$$

where $p(D) = \int p(D|t)p(t)dt$

- normalizing constant, not a function of t
- typically computationally expensive

We need to know the posterior distribution up to constant:

$$\underbrace{p(t \mid D)}_{\text{posterior distribution}} \quad \propto \quad \underbrace{p(D \mid t)}_{\text{Likeihood } L(t|D)} \quad \times \quad \underbrace{p(t)}_{\text{prior distribution}}$$

- Sample $t_1, \ldots, t_n$ from a function proportional to $p(D|t)p(t)$ approximates the posterior distribution $p(t|D)$
- Markov chain Monte Carlo simulation

Example: Bayesian inference

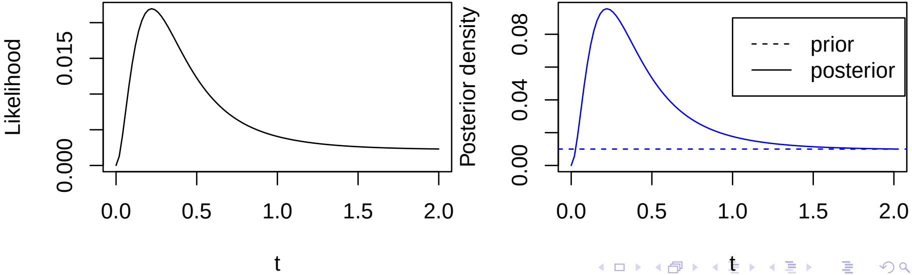
Likelihood:

$$L(t) = p(D|t) = \left(\frac{3}{4} - \frac{3}{4}e^{-8t/3}\right)^2 \left(\frac{1}{4} + \frac{3}{4}e^{-8t/3}\right)^4$$

Consider prior distribution: $t \sim U(0, 100)$

Then the posterior distribution of t:

$$p(t|D) \propto p(D|t)p(t)$$



Prior Distributions

- Informative vs. Uninformative priors
  - A prior distribution which is non-commital about a parameter, for example, a uniform distribution.
- Proper vs. Improper priors
  - An improper prior function integrates to infinity. Not a probability distribution. EX) $p(t) \propto 1$
  - An improper prior is usually accepted when the corresponding posterior distribution is proper.
- Conjugate prior
  - the prior and posterior have the same distributional form.
  - analytical computation
  - not useful in phylogenetic tree inference, because the model is typically too complex.

Example: Effects of Priors

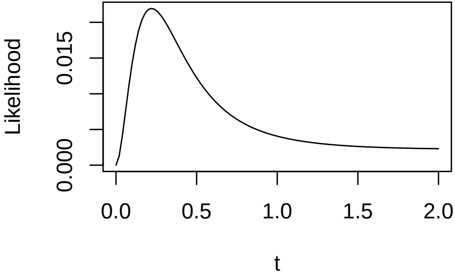
Consider two prior distributions:

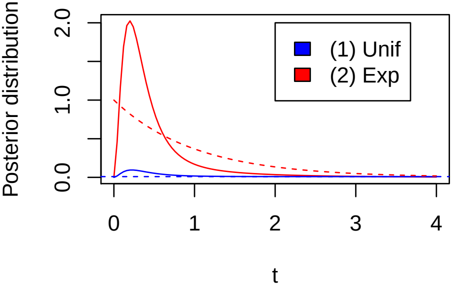
(1) $t \sim U(0, 100)$

- Noninformative prior

(2) $t \sim Exp(1)$

- Typically branch length < 1.





How to choose priors

- Subjective prior vs Objective prior
    - Expert's subjective belief: how to quantify?
    - Objective belief: A commonly used alternative. Typically uninformative prior or low-informative prior
- Posterior's sensitivity to prior
    - When data is very informative, the likelihood dominates the posterior. less sensitive to prior choice.
    - Important to assess the influence of the prior
      ex. Tune the parameters in prior. Consider $Exp(1)$ and $Exp(2)$ as priors. If the posterior results are quite different, you may consider less informative prior.

Monte Carlo Interation

A simulation method for calculating multidimensional integrals

Example 1. Want to compute the posterior mean $E(t|D) = \int tp(t|D)dt$.

If we have sample $t_1, \ldots, t_n \sim p(t|D)$, the estimate of the posterior mean is

$$\frac{1}{n}\sum_{i=1}^{n} t_i = \bar{t}$$

Monte Carlo Interation

A simulation method for calculating multidimensional integrals

Example 2. Consider multiple parameters $\theta_1, \ldots, \theta_k$. Want to compute the posterior mean for θ_1,

$$E(\theta_1|D) = \int \theta_1 p(\theta_1|D) d\theta_1,$$

where $p(\theta_1|D) = \int \cdots \int p(\theta_1, \theta_2, \ldots, \theta_k|D) d\theta_2 \cdots d\theta_k$. Let's say we sample parameters from their posterior: $\{\theta_{1,i}, \ldots, \theta_{k,i}\} \sim p(\theta_1, \ldots, \theta_k|D)$. Taking $\theta_{1,1}, \ldots, \theta_{1,n}$ out of the sample forms the marginal posterior distribution. Therefore, the estimate of the posterior mean is

$$\frac{1}{n}\sum_{i=1}^{n} \theta_{1,i}$$

A Markov chain Monte Carlo (MCMC) simulation

- An MCMC is a method that makes valid, but dependent, draws from the posterior distribution of interest
- A Markov chain: $t_0, t_1, t_2, \ldots$
    - state: t_i (sampled value)
    - state space: the space of parameters
        - ★ tree topology: all possible topologies
        - ★ branch lengths: any positive values
    - Markov Property

$$P(t_i | t_0, \ldots, t_{i-1}) = P(t_i | t_{i-1})$$

Metropolis-Hastings (MH) Algorithm

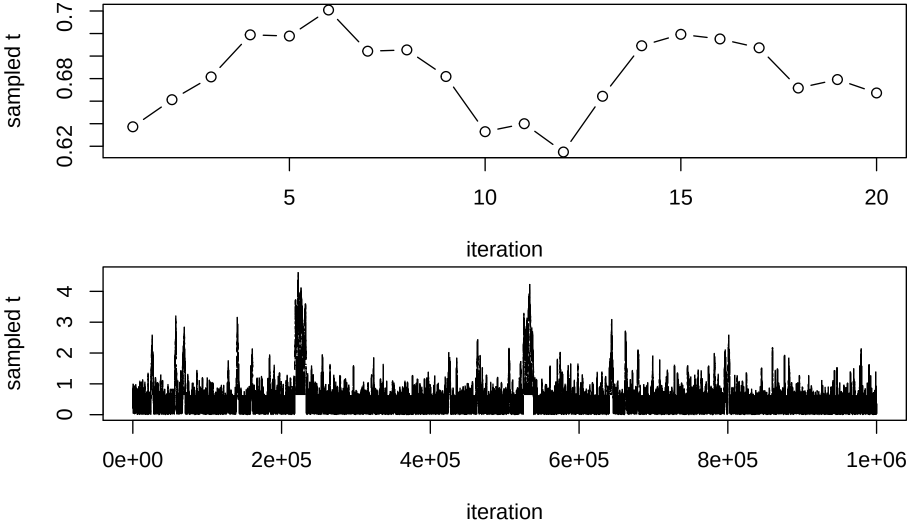
Example: two sequences

1. t_0 is the initial branch length. Let t_i be the state at the ith iteration.
2. At iteration $i+1$, a new state t^* is proposed from function $q(t^*|t_i)$. (ex. Uniform on $(t_i-\epsilon, t_i+\epsilon)$ for some small $\epsilon > 0$)
3. The proposed state t^* is accepted with probability

$$\min\left(1,\ \underbrace{\frac{L(t^*|D)}{L(t_i|D)}}_{\text{likelihood ratio}} \times \underbrace{\frac{p(t^*)}{p(t_i)}}_{\text{prior ratio}} \times \underbrace{\frac{q(t_i|t^*)}{q(t^*|t_i)}}_{\text{proposal ratio}}\right)$$

4. If t^* is accepted, $t_{i+1} = t^*$. Otherwise, $t_{i+1} = t_i$.
5. Repeat (2)-(4) until converge.

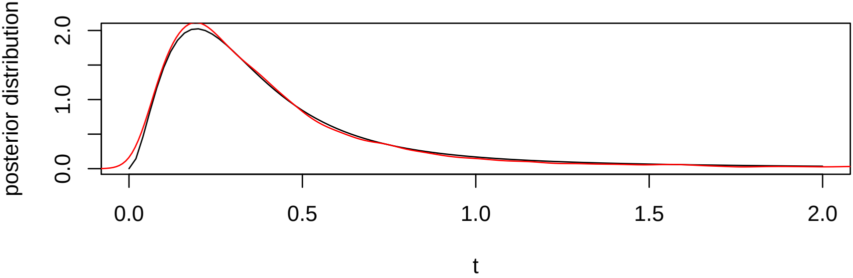
The algorithm generates a Markov chain



Metropolis-Hastings (MH) Algorithm

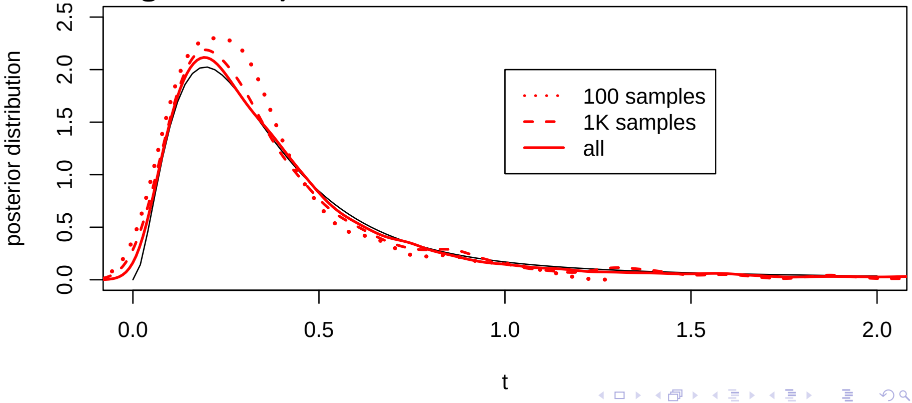
Saving the sample after

- burn-in: typically remove the first 10-30% of sample.
- thining: for example, save every 100th iteration, less correlated



Metropolis-Hastings (MH) Algorithm

If a proposal function satifies some regularity conditions (the Markov chain is aperiodic and irreducible), **the proportion of the time that the Markov chain visited a state approximates the posterior probability of the state. In other words, the empirical distribution of the sample will converge on the posterior distribution.**



MCMC for Phylogenetic tree estimation?

- Many parameters
  - tree topology, tree branch, substitution rate parameters etc
- able to simulate all parameters from their posterior distribution
- Easy to approximate the marignal posterior probability of tree topology from the joint samples.

Summarize the sample

- branch lengths (continuous): posterior mean
- tree topology (discrete)
  - the maximum a *posteriori* (MAP) tree: the one with highest posterior probability
  - majority-rule tree: combines the most common clades, and usually yields a tree that wasn't sampled in the analysis
  - maximum clade credibility tree: Each clade within the tree is given a score based on the fraction of times that it appears in the set of sampled posterior trees, and the product of these scores are taken as the tree's score. The tree with the highest score is then the maximum clade credibility tree

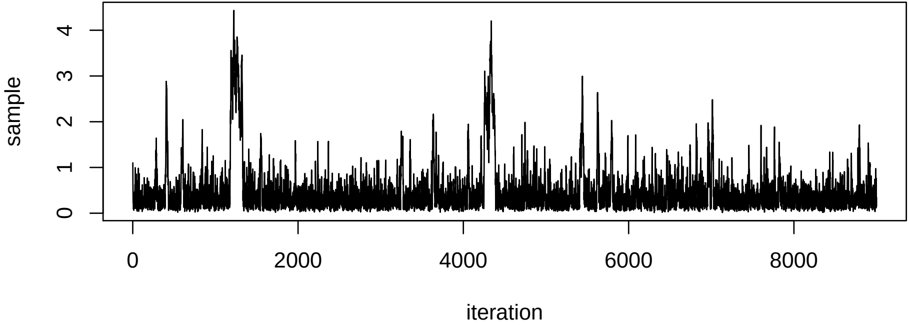
MCMC Diagnostic

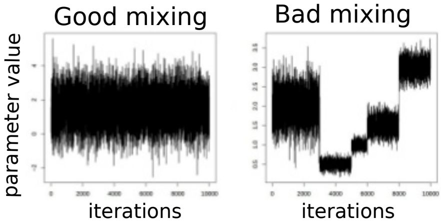
- How to know if the sample converges to the posterior distribution
- How to know if the Markov chain visited most plausible states (mixing)

Tools

- R pakcage `CODA`
- Some software provide diagnostic tools as well.

MCMC Diagnostic: trace plot





MCMC Diagnostic

- Acceptance rate: not too high, not too low (30-40%)
  - too high: slow convergence
  - too low: mixing problem
- Multiple independent runs: comparing the samples (distributions, posterior means)
- Statistics
  - effective sample size (ESS): higher, better
  - Gelman and Rubin's statistic or *Potential scale reduction Factor (PSRF)* (in `CODA`): the statistic < 1.1 or 1.2 indicates convergence
- Topology
  - comparing the distribution of bipartitions between two or more independent runs

Limitations of MCMC

- difficult to determine if the chain has converged to the desired distribution
- mixing problem: too large state space and correlated parameters
- expensive computing time.

Software

Phylogenetic gene tree

- MrBayes, BEAST

Species tree

- starBEAST, BEST
- BUCKy (concordance tree)

Divergence time

- IMa2
- Migrate