Bayesian Tree Sampling

Heiko Schmidt / Greg Ewing

June 7, 2007

The difference

The Bayesian approach asks the right question in a hypothesis testing procedure, namely, "What is the probability that this hypothesis is true, given the data?" rather than the classical approach, which asks a question like, "Assuming that this hypothesis is true, what is the probability of the observed data?"

–Statistical Methods in Bioinformatics

Derivation

We know that

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A),$$

from conditional probability.

Derivation

We know that

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A),$$

from conditional probability. Also

$$\Pr(A \cap B) = \Pr(B \cap A) = \Pr(A|B)\Pr(B).$$

Derivation

We know that

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A),$$

from conditional probability. Also

$$\Pr(A \cap B) = \Pr(B \cap A) = \Pr(A|B)\Pr(B).$$

Therefore

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$
$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

This is Bayes formula or theorem.

Bayes Theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

- Bayesian, flips the probability around.

Bayes Theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

$$\underbrace{\Pr(A|B)}_{\text{Posterior Density}} \propto \overbrace{\mathrm{L}(A,B)}^{\text{Likelihood}}\overbrace{\Pr(A)}^{\text{Prior}}$$

- Bayesian, flips the probability around.
- It is easy to include prior information which is often available.

Bayes Theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

$$\underbrace{\Pr(A|B)}_{\text{Posterior Density}} \propto \overbrace{\mathrm{L}(A,B)}^{\text{Likelihood}}\overbrace{\Pr(A)}^{\text{Prior}}$$

- Bayesian, flips the probability around.
- It is easy to include prior information which is often available.
- The Bayesian conditional probability is perhaps more intuitive.

Making formulas tangible

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $L(T, D, M) = \Pr(D|T, M)$

Making formulas tangible

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $L(T, D, M) = \Pr(D|T, M)$

  - T is the tree.
  - D is the DNA/Protein etc sequence data.
  - M is the model parameters, like GTR.

Making formulas tangible

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $L(T, D, M) = \Pr(D|T, M)$
  - T is the tree.
  - D is the DNA/Protein etc sequence data.
  - M is the model parameters, like GTR.

- In words: The likelihood is the probability of the DNA data given the Tree and the model parameters.

Making formulas tangible

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $L(T, D, M) = \Pr(D|T, M)$
  - T is the tree.
  - D is the DNA/Protein etc sequence data.
  - M is the model parameters, like GTR.
- In words: The likelihood is the probability of the DNA data given the Tree and the model parameters.

- The Prior is $\Pr(T, M)$ and indicates any information we already know. E.g., the root is not older than 10 million years.

Making formulas tangible

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $L(T, D, M) = \Pr(D|T, M)$
  - T is the tree.
  - D is the DNA/Protein etc sequence data.
  - M is the model parameters, like GTR.
- In words: The likelihood is the probability of the DNA data given the Tree and the model parameters.
- The Prior is $\Pr(T, M)$ and indicates any information we already know. E.g., the root is not older than 10 million years.

- The Posterior density is $\Pr(T, M|D)$ the probability of the tree and model parameters given the sequence data.

The Bad News

- We can't directly solve for the posterior distribution.

The Bad News

- We can't directly solve for the posterior distribution.

- Therefore MHMCMC must be used, this means it will take a lot of computer resources.

The Bad News

- We can't directly solve for the posterior distribution.
- Therefore MHMCMC must be used, this means it will take a lot of computer resources.

- The "answer" is not a tree, but a distribution of trees/states.

The Bad News

- We can't directly solve for the posterior distribution.
- Therefore MHMCMC must be used, this means it will take a lot of computer resources.
- The “answer” is not a tree, but a distribution of trees/states.
- It will always be slower than ML.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.
- Example: Our machine flips a coin and either adds one to the last output or subtracts one.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.
- Example: Our machine flips a coin and either adds one to the last output or subtracts one.

Machine Output

1,2,1,0,1,0,-1,-2,-3,-2,-3,-4,-3,-2,-1,0,-1,0,1,2,1,2,3

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.
- Example: Our machine flips a coin and either adds one to the last output or subtracts one.

Machine Output

1,2,1,0,1,0,-1,-2,-3,-2,-3,-4,-3,-2,-1,0,-1,0,1,2,1,2,3

- We don't care about the whole sequence, just the last output which is 3.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.
- Example: Our machine flips a coin and either adds one to the last output or subtracts one.

Machine Output

1,2,1,0,1,0,-1,-2,-3,-2,-3,-4,-3,-2,-1,0,-1,0,1,2,1,2,3

- We don't care about the whole sequence, just the last output which is 3.
- The next item has a 50% chance that it will be a 4, and a 50% chance that it will be a 2.

Markov Chains

- Assume that I have a machine that outputs random numbers, ie a chain of numbers.
- If I can work out the probability of the next output by only looking at the previous output, it is said to have the Markov property.
- Example: Our machine flips a coin and either adds one to the last output or subtracts one.

Machine Output

1,2,1,0,1,0,-1,-2,-3,-2,-3,-4,-3,-2,-1,0,-1,0,1,2,1,2,3

- We don't care about the whole sequence, just the last output which is 3.
- The next item has a 50% chance that it will be a 4, and a 50% chance that it will be a 2.
- This is a Markov Chain.

Definition of a Markov Chain

Definition

A Markov Chain is a chain of randomly chosen values where the probability of the next value is entirely determined by the previous value.

Definition of a Markov Chain

Definition

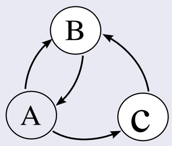
A Markov Chain is a chain of randomly chosen values where the probability of the next value is entirely determined by the previous value.

Rough Math definition

$$\Pr(X_n | X_{n-1}, X_{n-2}, \ldots) = \Pr(X_n | X_{n-1})$$

Markov Chain Graph

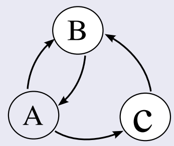
State Graph



- Simple Markov Chains can be represented as a graph.

Markov Chain Graph

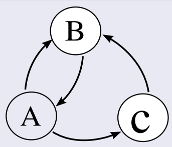
State Graph



- Simple Markov Chains can be represented as a graph.
- Nodes or circles represent states (the last output).

Markov Chain Graph

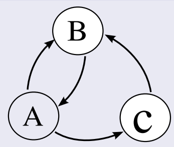
State Graph



- Simple Markov Chains can be represented as a graph.
- Nodes or circles represent states (the last output).
- Arrows are transitions between states.

Markov Chain Graph

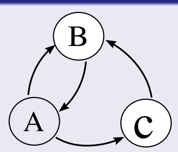
State Graph



- Simple Markov Chains can be represented as a graph.
- Nodes or circles represent states (the last output).
- Arrows are transitions between states.
- Transitions (Arrows) usually have probabilities on them. That is the probability that this transition will be followed.

Markov Chain Graph

State Graph

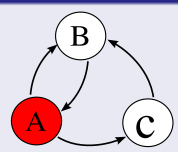


- Simple Markov Chains can be represented as a graph.
- Nodes or circles represent states (the last output).
- Arrows are transitions between states.
- Transitions (Arrows) usually have probabilities on them. That is the probability that this transition will be followed.
- For clarity, when transitions are equiprobable we omit the transition probabilities.

Markov Chain Graph

Example

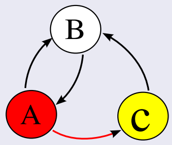
State Graph



Output

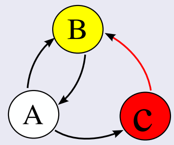
A

Markov Chain Graph
Example

State Graph


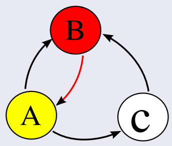
Output
A C

Markov Chain Graph
Example

State Graph


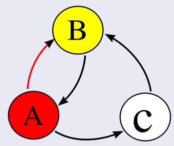
Output
A C B

Markov Chain Graph
Example

State Graph


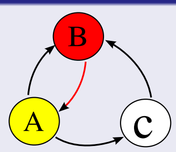
Output
A C B A

Markov Chain Graph
Example

State Graph


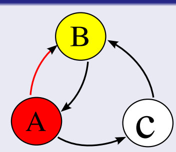
Output
A C B A B

Markov Chain Graph
Example

State Graph


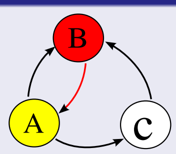
Output
A C B A B A

Markov Chain Graph
Example

State Graph


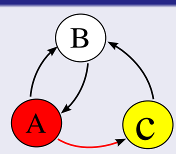
Output
A C B A B A B

Markov Chain Graph
Example

State Graph


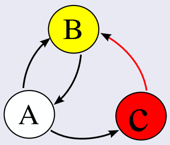
Output
A C B A B A B A

Markov Chain Graph
Example

State Graph


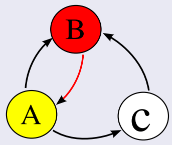
Output
A C B A B A B A C

Markov Chain Graph
Example

State Graph


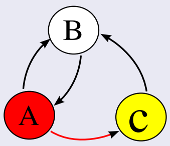
Output
A C B A B A B A C B

Markov Chain Graph
Example

State Graph


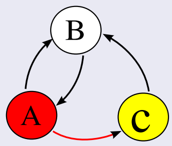
Output
A C B A B A B A C B A

Markov Chain Graph
Example

State Graph


Output
A C B A B A B A C B A C

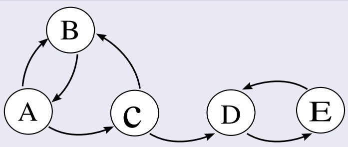
Markov Chain Graph
Example

State Graph


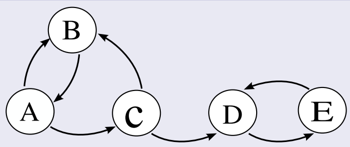
Output
A C B A B A B A C B A C

- Note that the states can be anything. ie different trees

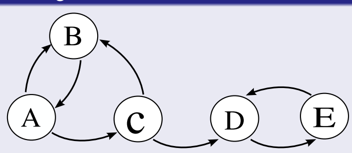
Extra Markov Chain Properties
Irreducibility

Reducible state diagram


Extra Markov Chain Properties
Irreducibility

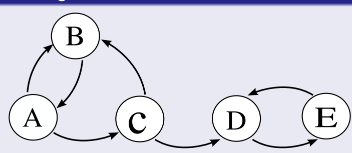
Reducible state diagram


Extra Markov Chain Properties
Irreducibility

Reducible state diagram


Definition
A Markov Chain is Irreducible if and only if the chain can get from any possible state to any other possible state eventually.

Extra Markov Chain Properties
Irreducibility

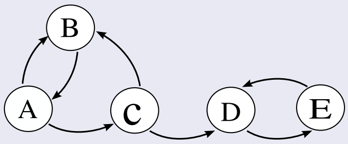
Reducible state diagram


Definition
A Markov Chain is Irreducible if and only if the chain can get from any possible state to any other possible state eventually.

- The above state diagram is NOT irreducible.

Extra Markov Chain Properties
Irreducibility

Reducible state diagram

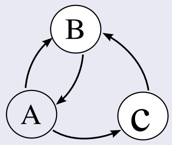


Definition

A Markov Chain is Irreducible if and only if the chain can get from any possible state to any other possible state eventually.

- The above state diagram is NOT irreducible.
- Adding a transition from $D \rightarrow C$ it would make this irreducible

Extra Markov Chain Properties
Reversibility

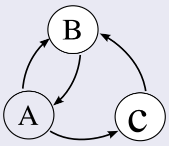


Is this output reversed?

C A B C A B A B A B C

- Note that there is no $C \rightarrow B$ transition or $C \rightarrow A$ transition.

Extra Markov Chain Properties
Reversibility



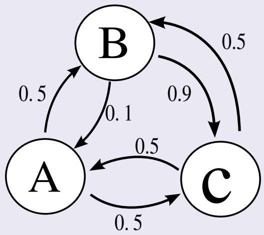
Is this output reversed?

C A B C A B A B A B C

- Note that there is no $C \rightarrow B$ transition or $C \rightarrow A$ transition.
- Therefore we can tell that this output sequence is reversed.

Extra Markov Chain Properties
Reversibility

Tricky Example



Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

Extra Markov Chain Properties
Reversibility

Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

- The transition $B \rightarrow A$ is much less likely than $B \rightarrow C$ in the forward direction.

Extra Markov Chain Properties
Reversibility

Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

- The transition $B \rightarrow A$ is much less likely than $B \rightarrow C$ in the forward direction.
- In this example there are 7 $B \rightarrow C$ transitions and only 1 $B \rightarrow A$ transition in the forward direction.

Extra Markov Chain Properties
Reversibility

Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

- The transition $B \rightarrow A$ is much less likely than $B \rightarrow C$ in the forward direction.
- In this example there are 7 $B \rightarrow C$ transitions and only 1 $B \rightarrow A$ transition in the forward direction.
- Conversely there are 4 $B \rightarrow C$ transitions and 4 $B \rightarrow A$ transitions in the reverse direction.

Extra Markov Chain Properties
Reversibility

Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

- The transition $B \rightarrow A$ is much less likely than $B \rightarrow C$ in the forward direction.
- In this example there are 7 $B \rightarrow C$ transitions and only 1 $B \rightarrow A$ transition in the forward direction.
- Conversely there are 4 $B \rightarrow C$ transitions and 4 $B \rightarrow A$ transitions in the reverse direction.
- It seems we can guess that this output is not reversed.

Extra Markov Chain Properties
Reversibility

Is this output reversed?

A B C A B C B C B C B C A B C B C A B A

- The transition $B \to A$ is much less likely than $B \to C$ in the forward direction.
- In this example there are 7 $B \to C$ transitions and only 1 $B \to A$ transition in the forward direction.
- Conversely there are 4 $B \to C$ transitions and 4 $B \to A$ transitions in the reverse direction.
- It seems we can guess that this output is not reversed.
- But we stick to simple definitions for this workshop.

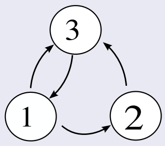
Extra Markov Chain Properties
Reversibility

Definition

A Markov Chain is reversible if we cannot detect whether or not the chain is running in “reverse”. That is the output is statistically identicle in both directions.

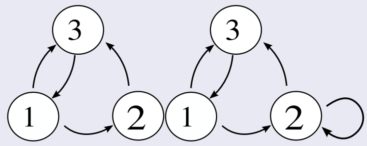
Extra Markov Chain Properties
Aperiodic

Periodic-Aperiodic



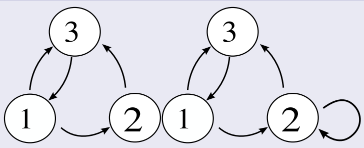
Extra Markov Chain Properties
Aperiodic

Periodic-Aperiodic



Extra Markov Chain Properties
Aperiodic

Periodic-Aperiodic



Definition

A Markov Chain is periodic if there is some fixed “cycle” of states, and it is aperiodic otherwise.

Extra Markov Chain Properties

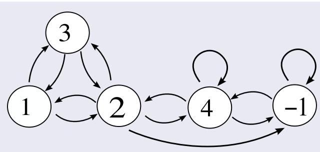
Why do we care?

Extra Markov Chain Properties

Why do we care?

- If a MCMC chain has these 3 properties (reversible, irreducible and aperiodic), then it is also ergodic.

Extra Markov Chain Properties
Stationary distribution

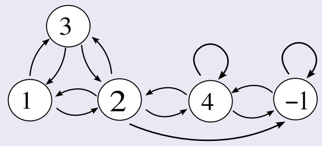


output

1 3 2 4 4 2 1 2 -1 -1 4 2 3 1 3 2 4 4 4 -1 -1 -1 4 2 3 1 2 3 -1

- We can calculate statistics on the output, like mean and standard deviation. Also we can plot histograms etc.

Extra Markov Chain Properties
Stationary distribution

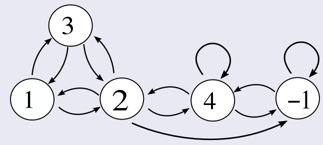


output

1 3 2 4 4 2 1 2 -1 -1 4 2 3 1 3 2 4 4 4 -1 -1 -1 4 2 3 1 2 3 -1

- We can calculate statistics on the output, like mean and standard deviation. Also we can plot histograms etc.
- Consider the distribution of the output.

Extra Markov Chain Properties
Stationary distribution



output

1 3 2 4 4 2 1 2 -1 -1 4 2 3 1 3 2 4 4 4 -1 -1 -1 4 2 3 1 2 3 -1

- We can calculate statistics on the output, like mean and standard deviation. Also we can plot histograms etc.
- Consider the distribution of the output.
- What about the start state. That is if the chain is started in state 1, will the distribution be different from starting in 2.

Extra Markov Chain Properties
Ergodic

Definition

If we can start from any state, and if we take samples for long enough, and we end up with the same distribution, that distribution is the stationary distribution of the Markov Chain, and the Markov Chain is said to be ergodic

Extra Markov Chain Properties
Ergodic

Definition

If we can start from any state, and if we take samples for long enough, and we end up with the same distribution, that distribution is the stationary distribution of the Markov Chain, and the Markov Chain is said to be ergodic

Definition

If a Markov Chain is reversible, irreducible and aperiodic then it is also ergodic.

Extra Markov Chain Properties
Ergodic

Definition

If we can start from any state, and if we take samples for long enough, and we end up with the same distribution, that distribution is the stationary distribution of the Markov Chain, and the Markov Chain is said to be ergodic

Definition

If a Markov Chain is reversible, irreducible and aperiodic then it is also ergodic.

- So we can know that a chain will converge to the stationary distribution without testing every state.

Extra Markov Chain Properties
Ergodic

Definition

If we can start from any state, and if we take samples for long enough, and we end up with the same distribution, that distribution is the stationary distribution of the Markov Chain, and the Markov Chain is said to be ergodic

Definition

If a Markov Chain is reversible, irreducible and aperiodic then it is also ergodic.

- So we can know that a chain will converge to the stationary distribution without testing every state.
- Usually the symbol π denotes the stationary distribution.

Extra Markov Chain Properties
Ergodic

Definition

If we can start from any state, and if we take samples for long enough, and we end up with the same distribution, that distribution is the stationary distribution of the Markov Chain, and the Markov Chain is said to be ergodic

Definition

If a Markov Chain is reversible, irreducible and aperiodic then it is also ergodic.

- So we can know that a chain will converge to the stationary distribution without testing every state.
- Usually the symbol π denotes the stationary distribution.
- Note that we have not said anything about how many samples we need to get an accurate distribution.

Metropolis Hastings MCMC

Algorithm

- Start in state X_n

Metropolis Hastings MCMC

Algorithm

- Start in state X_n
- Randomly generate some new state X' from X

Metropolis Hastings MCMC

Algorithm

- Start in state X_n
- Randomly generate some new state X' from X
- Calculate the acceptance probability based on the posterior density.

Metropolis Hastings MCMC

Algorithm

- Start in state X_n
- Randomly generate some new state X' from X
- Calculate the acceptance probability based on the posterior density.
- Accept the new state with that probability.

Metropolis Hastings MCMC

Algorithm

- Start in state X_n
- Randomly generate some new state X' from X
- Calculate the acceptance probability based on the posterior density.
- Accept the new state with that probability.
- If we accept, then $X_{n+1} = X'$, otherwise $X_{n+1} = X_n$.

- If our new state generation step can get to any valid state eventually (with non zero probability), then the chain is irreducible.

- If our new state generation step can get to any valid state eventually (with non zero probability), then the chain is irreducible.
- If it's possible to generate X' from X and X from X' then the chain can be reversible.

- If our new state generation step can get to any valid state eventually (with non zero probability), then the chain is irreducible.
- If it's possible to generate X' from X and X from X' then the chain can be reversible.
- The acceptance probability is chosen so that the chain will be reversible and aperiodic.

- If our new state generation step can get to any valid state eventually (with non zero probability), then the chain is irreducible.
- If it's possible to generate X' from X and X from X' then the chain can be reversible.
- The acceptance probability is chosen so that the chain will be reversible and aperiodic.
- Therefore the chain is ergodic with stationary distribution π.

- If our new state generation step can get to any valid state eventually (with non zero probability), then the chain is irreducible.
- If it's possible to generate X' from X and X from X' then the chain can be reversible.
- The acceptance probability is chosen so that the chain will be reversible and aperiodic.
- Therefore the chain is ergodic with stationary distribution π.

The Key Idea

The stationary distribution is the posterior distribution of interest. That is the MHMCMC chain is sampling the Bayesian posterior distribution.

Example

- Start with tree $T = (a, b|c, d)$.

Output

$(a, b|c, d)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$

Output

$(a, b|c, d)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$
- Calculate acceptance probability and then accept/reject. We reject this time.

Output

$(a, b|c, d)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$
- Calculate acceptance probability and then accept/reject. We reject this time.
- The new state is $T = (a, b|c, d)$ which we output.

Output

$(a, b|c, d)$ $(a, b|c, d)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$
- Calculate acceptance probability and then accept/reject. We reject this time.
- The new state is $T = (a, b|c, d)$ which we output.
- The next generated state is $T' = (a, d|b, c)$ ($b \rightleftharpoons d$) and this time we accept.

Output

$(a, b|c, d)$ $(a, b|c, d)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$
- Calculate acceptance probability and then accept/reject. We reject this time.
- The new state is $T = (a, b|c, d)$ which we output.
- The next generated state is $T' = (a, d|b, c)$ ($b \rightleftharpoons d$) and this time we accept.
- The new state is $T = (a, d|b, c)$

Output

$(a, b|c, d)$ $(a, b|c, d)$ $(a, d|b, c)$

Example

- Start with tree $T = (a, b|c, d)$.
- Generate a new tree from T by a branch swap ($b \rightleftharpoons c$). $T' = (a, c|b, d)$
- Calculate acceptance probability and then accept/reject. We reject this time.
- The new state is $T = (a, b|c, d)$ which we output.
- The next generated state is $T' = (a, d|b, c)$ ($b \rightleftharpoons d$) and this time we accept.
- The new state is $T = (a, d|b, c)$
- We continue $T' = (a, c|b, d)$ ($c \rightleftharpoons d$), and accept.

Output

$(a, b|c, d)$ $(a, b|c, d)$ $(a, d|b, c)$ $(a, c|b, d)$

Die example

Wiki Formula

$$\Pr(k|i,s) = \frac{1}{s^i}\sum_{n=0}^{\lfloor \frac{k-i}{s} \rfloor} (-1)^n \binom{i}{n}\binom{k-sn-1}{i-1}$$

Die MHMCMC

- Formula looks too complicated!

Die example

Wiki Formula

$$\Pr(k|i,s) = \frac{1}{s^i}\sum_{n=0}^{\lfloor \frac{k-i}{s} \rfloor} (-1)^n \binom{i}{n}\binom{k-sn-1}{i-1}$$

Die MHMCMC

- Formula looks too complicated!
- Use a simple MHMCMC instead.

Die example

Wiki Formula

$$\Pr(k|i,s) = \frac{1}{s^i}\sum_{n=0}^{\lfloor \frac{k-i}{s} \rfloor} (-1)^n \binom{i}{n}\binom{k-sn-1}{i-1}$$

Die MHMCMC

- Formula looks too complicated!
- Use a simple MHMCMC instead.
- Just pick one die at random and re-throw.

Die example

Wiki Formula

$$\Pr(k|i,s) = \frac{1}{s^i}\sum_{n=0}^{\lfloor \frac{k-i}{s} \rfloor} (-1)^n \binom{i}{n}\binom{k-sn-1}{i-1}$$

Die MHMCMC

- Formula looks too complicated!
- Use a simple MHMCMC instead.
- Just pick one die at random and re-throw.
- This is reversible and the acceptance ratio is 1. i.e we always accept.

Die example

3 die

1	1	1

Output

3

Die example

3 die

1	1	1
4	1	1

Output

3 6

Die example

3 die

1	1	1
4	1	1
4	1	6

Output

3 6 11

Die example

3 die

1	1	1
4	1	1
4	1	6
2	1	6

Output

3 6 11 9

Die example

3 die

1	1	1
4	1	1
4	1	6
2	1	6
3	1	6

Output

3 6 11 9 10

Die example

3 die

1	1	1
4	1	1
4	1	6
2	1	6
3	1	6
3	1	4

Output

3 6 11 9 10 8

Die example

3 die

1	1	1
4	1	1
4	1	6
2	1	6
3	1	6
3	1	4
3	5	4

Output

3 6 11 9 10 8 12

Die example

3 die

1	1	1
4	1	1
4	1	6
2	1	6
3	1	6
3	1	4
3	5	4
3	2	4

Output

3 6 11 9 10 8 12 9

More Die

- By changing just one dice at each step, the sum can never change by more than 5 from step to step.

More Die

- By changing just one dice at each step, the sum can never change by more than 5 from step to step.
- If we have 100 die and start at all ones, it will take a long time to get to the “equilibrium”.

More Die

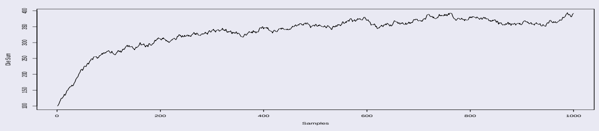
- By changing just one dice at each step, the sum can never change by more than 5 from step to step.
- If we have 100 die and start at all ones, it will take a long time to get to the “equilibrium”.
- On the other hand we could roll *every die at each step.*

More Die

- By changing just one dice at each step, the sum can never change by more than 5 from step to step.
- If we have 100 die and start at all ones, it will take a long time to get to the “equilibrium”.
- On the other hand we could roll *every die at each step.*
- In this case we get to equilibrium in just a single step but must generate 100 random numbers per step.

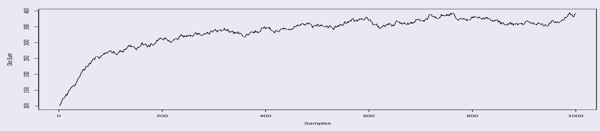
More Die

100 die, rolling 1 dice per step

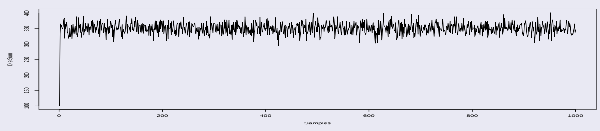


More Die

100 die, rolling 1 dice per step



100 die, rolling all per step



Effective Sample Size

- Both chains were 1000 MCMC samples long, but each sample is not independent of the other.

Effective Sample Size

- Both chains were 1000 MCMC samples long, but each sample is not independent of the other.
- Its clear that the second case gives better results.

Effective Sample Size

- Both chains were 1000 MCMC samples long, but each sample is not independent of the other.
- Its clear that the second case gives better results.
- Effective sample size is the estimated number of independent samples and is calculated with the Integrated autocorrelation time. (in tracer for example)

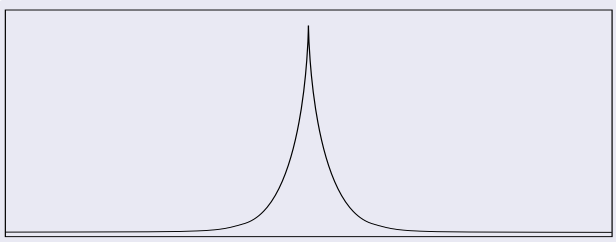
Effective Sample Size

- Both chains were 1000 MCMC samples long, but each sample is not independent of the other.
- Its clear that the second case gives better results.
- Effective sample size is the estimated number of independent samples and is calculated with the Integrated autocorrelation time. (in tracer for example)
- Due to the correlations between samples we don't really need every sample from the MCMC chain and instead only collect every 100'th sample or so.

Effective Sample Size

- Both chains were 1000 MCMC samples long, but each sample is not independent of the other.
- Its clear that the second case gives better results.
- Effective sample size is the estimated number of independent samples and is calculated with the Integrated autocorrelation time. (in tracer for example)
- Due to the correlations between samples we don't really need every sample from the MCMC chain and instead only collect every 100'th sample or so.
- Performance should be measured in the number of effective samples per CPU cycle.

Witch's Hat



Witch's Hat

- Consider all non tree-like signals.
- Recombination, Horizontal Gene Transfer and other effects could contribute to a lot of witch's hats.

Key Points for simple analysis

- Check Effective Sample Size.

Key Points for simple analysis

- Check Effective Sample Size.
- Choose the correct sample intervals.

Key Points for simple analysis

- Check Effective Sample Size.
- Choose the correct sample intervals.
- Check Burn-in. It should be small enough that it does not matter if you include it.

Key Points for simple analysis

- Check Effective Sample Size.
- Choose the correct sample intervals.
- Check Burn-in. It should be small enough that it does not matter if you include it.
- Not all moves are equal. How long depends on many things

Key Points for simple analysis

- Check Effective Sample Size.
- Choose the correct sample intervals.
- Check Burn-in. It should be small enough that it does not matter if you include it.
- Not all moves are equal. How long depends on many things
- Multiple runs from random starting locations

Posterior

$$\Pr(T, M|D) \propto \Pr(D|T, M)\Pr(T, M)$$

- The likelihood is $\mathrm{L}(T, D, M) = \Pr(D|T, M)$
- T is the tree.
- D is the DNA/Protein etc sequence data.
- M is the model parameters, like GTR.

Warning

Trees Make Life Difficult

Moves and why you care about irreducibility

- Many programs have a huge set of options.

Moves and why you care about irreducibility

- Many programs have a huge set of options.
- It is often possible to have moves that are not reversible or irreducible.

Moves and why you care about irreducibility

- Many programs have a huge set of options.
- It is often possible to have moves that are not reversible or irreducible.
- Hence will not properly sample the posterior distribution.

Moves and why you care about irreducibility

- Many programs have a huge set of options.
- It is often possible to have moves that are not reversible or irreducible.
- Hence will not properly sample the posterior distribution.
- It may not be possible to get to the parts of the state space that are of interest.

Moves and why you care about irreducibility

- Many programs have a huge set of options.
- It is often possible to have moves that are not reversible or irreducible.
- Hence will not properly sample the posterior distribution.
- It may not be possible to get to the parts of the state space that are of interest.
- The wrong choice of moves could make the chain run very slowly.

Moves and why you care about irreducibility

- Many programs have a huge set of options.
- It is often possible to have moves that are not reversible or irreducible.
- Hence will not properly sample the posterior distribution.
- It may not be possible to get to the parts of the state space that are of interest.
- The wrong choice of moves could make the chain run very slowly.
- Examples of real output.

Aside: Hot and Cold chains

- Have more than one chain.

Aside: Hot and Cold chains

- Have more than one chain.
- Each extra chain is heated. With only one chain that is not.

Aside: Hot and Cold chains

- Have more than one chain.
- Each extra chain is heated. With only one chain that is not.
- We swap states between chains at each step or as frequently as desired.

Aside: Hot and Cold chains

- Have more than one chain.
- Each extra chain is heated. With only one chain that is not.
- We swap states between chains at each step or as frequently as desired.
- Only collect samples from the cold chain. I.e., the only chain with the correct distribution.

Aside: Hot and Cold chains

- Have more than one chain.
- Each extra chain is heated. With only one chain that is not.
- We swap states between chains at each step or as frequently as desired.
- Only collect samples from the cold chain. I.e., the only chain with the correct distribution.
- The idea is that we won't get stuck.

Aside: Hot and Cold chains

- Have more than one chain.
- Each extra chain is heated. With only one chain that is not.
- We swap states between chains at each step or as frequently as desired.
- Only collect samples from the cold chain. I.e., the only chain with the correct distribution.
- The idea is that we won't get stuck.
- Generally not as effective as just developing some better moves.

Priors

- Huge topic!

Priors

- Huge topic!
- Without proper priors, the posterior density may not even exist!

Priors

- Huge topic!
- Without proper priors, the posterior density may not even exist!
- Priors do not need to be highly informed to be effective. e.g root height.

Priors

- Huge topic!
- Without proper priors, the posterior density may not even exist!
- Priors do not need to be highly informed to be effective. e.g root height.
- Informative priors can make analysis possible by restricting the state space

Priors

- Huge topic!
- Without proper priors, the posterior density may not even exist!
- Priors do not need to be highly informed to be effective. e.g root height.
- Informative priors can make analysis possible by restricting the state space
- Priors should be considered with respect to the hypothesis that will be tested.

Priors-Rules

- Trees must have a prior.

Priors-Rules

- Trees must have a prior.
- Even if all the branch lengths in a topology are infinitely long the likelihood is still finite.

Priors-Rules

- Trees must have a prior.
- Even if all the branch lengths in a topology are infinitely long the likelihood is still finite.
- Infinitely long branches do not make sense.

Priors-Rules

- Trees must have a prior.
- Even if all the branch lengths in a topology are infinitely long the likelihood is still finite.
- Infinitely long branches do not make sense.
- Yule priors, coalescent priors and exponential priors are common.

Priors-Rules

- Trees must have a prior.
- Even if all the branch lengths in a topology are infinitely long the likelihood is still finite.
- Infinitely long branches do not make sense.
- Yule priors, coalescent priors and exponential priors are common.
- For rooted topologies, a simple bounded uniform prior is sufficient.

Priors-Rules

- Trees must have a prior.
- Even if all the branch lengths in a topology are infinitely long the likelihood is still finite.
- Infinitely long branches do not make sense.
- Yule priors, coalescent priors and exponential priors are common.
- For rooted topologies, a simple bounded uniform prior is sufficient.
- Even if the max root height is 100 expected substitutions per site, the posterior can now be normalized.

Rules of thumb

- Do more than one run. I recommend about 10 or so if possible.

Rules of thumb

- Do more than one run. I recommend about 10 or so if possible.
- Each run should always start from a random starting point. Never use an NJ tree or any other “good” starting point.

Rules of thumb

- Do more than one run. I recommend about 10 or so if possible.
- Each run should always start from a random starting point. Never use an NJ tree or any other “good” starting point.
- Burn-in should be less than a tenth of the full run. In general if the statistics are affected by the amount of burn-in, it wasn’t run long enough.

Rules of thumb

- Do more than one run. I recommend about 10 or so if possible.
- Each run should always start from a random starting point. Never use an NJ tree or any other “good” starting point.
- Burn-in should be less than a tenth of the full run. In general if the statistics are affected by the amount of burn-in, it wasn’t run long enough.
- More parameters will always take longer. Don’t use more parameters than are needed.

Rules of thumb

- Do more than one run. I recommend about 10 or so if possible.
- Each run should always start from a random starting point. Never use an NJ tree or any other “good” starting point.
- Burn-in should be less than a tenth of the full run. In general if the statistics are affected by the amount of burn-in, it wasn’t run long enough.
- More parameters will always take longer. Don’t use more parameters than are needed.
- Check your priors!

Summary

- Bayesian inference is not maximum likelihood.

- Other points to consider.

Summary

- Bayesian inference is not maximum likelihood.
- It is not a black box. Care must be taken to get the chain setup correctly, and when interpreting the results.

- Other points to consider.

Summary

- Bayesian inference is not maximum likelihood.
- It is not a black box. Care must be taken to get the chain setup correctly, and when interpreting the results.
- Run the chain long enough! This is the most common mistake.

- Other points to consider.

Summary

- Bayesian inference is not maximum likelihood.
- It is not a black box. Care must be taken to get the chain setup correctly, and when interpreting the results.
- Run the chain long enough! This is the most common mistake.

- Other points to consider.
  - Generally slower than ML. (bootstrapped)

Summary

- Bayesian inference is not maximum likelihood.
- It is not a black box. Care must be taken to get the chain setup correctly, and when interpreting the results.
- Run the chain long enough! This is the most common mistake.

- Other points to consider.
  - Generally slower than ML. (bootstrapped)
  - Support values are easier to interpret.

Summary

- Bayesian inference is not maximum likelihood.
- It is not a black box. Care must be taken to get the chain setup correctly, and when interpreting the results.
- Run the chain long enough! This is the most common mistake.

- Other points to consider.
  - Generally slower than ML. (bootstrapped)
  - Support values are easier to interpret.
  - Can incorporate prior information easily.