Multitree Generalized Steppingstone Sampling – A New MCMC Method for Estimating the Marginal Likelihood of a Models

Mark T. Holder, Paul O. Lewis, David L. Swofford, and David Bryant

KU, UConn, Duke, U. Otago (NZ)

Feb 16, 2013 – Austin, TX

Context

- We rarely know the "true" model to use for analyses.
- Model-averaging methods can be difficult to implement and use.
- We can use the marginal likelihood to choose between models.

The marginal likelihood is the denominator of Bayes' Rule

$$p(\theta \mid D, M) \quad = \quad \frac{\mathbb{P}(D \mid \theta, M) p(\theta \mid M)}{\mathbb{P}(D \mid M)}$$

θ is a set of parameter values in the model M

D is the data

$\mathbb{P}(D \mid \theta, M)$ is the likelihood of θ.

$p(\theta, M)$ is the prior of θ.

The marginal likelihood assess the fit of the model to the data by considering all parameter values

$$\mathbb{P}(D \mid M) = \int \mathbb{P}(D \mid \theta, M) p(\theta \mid M) d\theta$$

MCMC avoids the marginal like. calc.

$$\frac{p(\theta^*|D,M)}{p(\theta \mid D,M)} = \frac{\frac{\mathbb{P}(D|\theta^*,M)p(\theta^*|M)}{\mathbb{P}(D|M)}}{\frac{\mathbb{P}(D|\theta,M)p(\theta|M)}{\mathbb{P}(D|M)}}$$

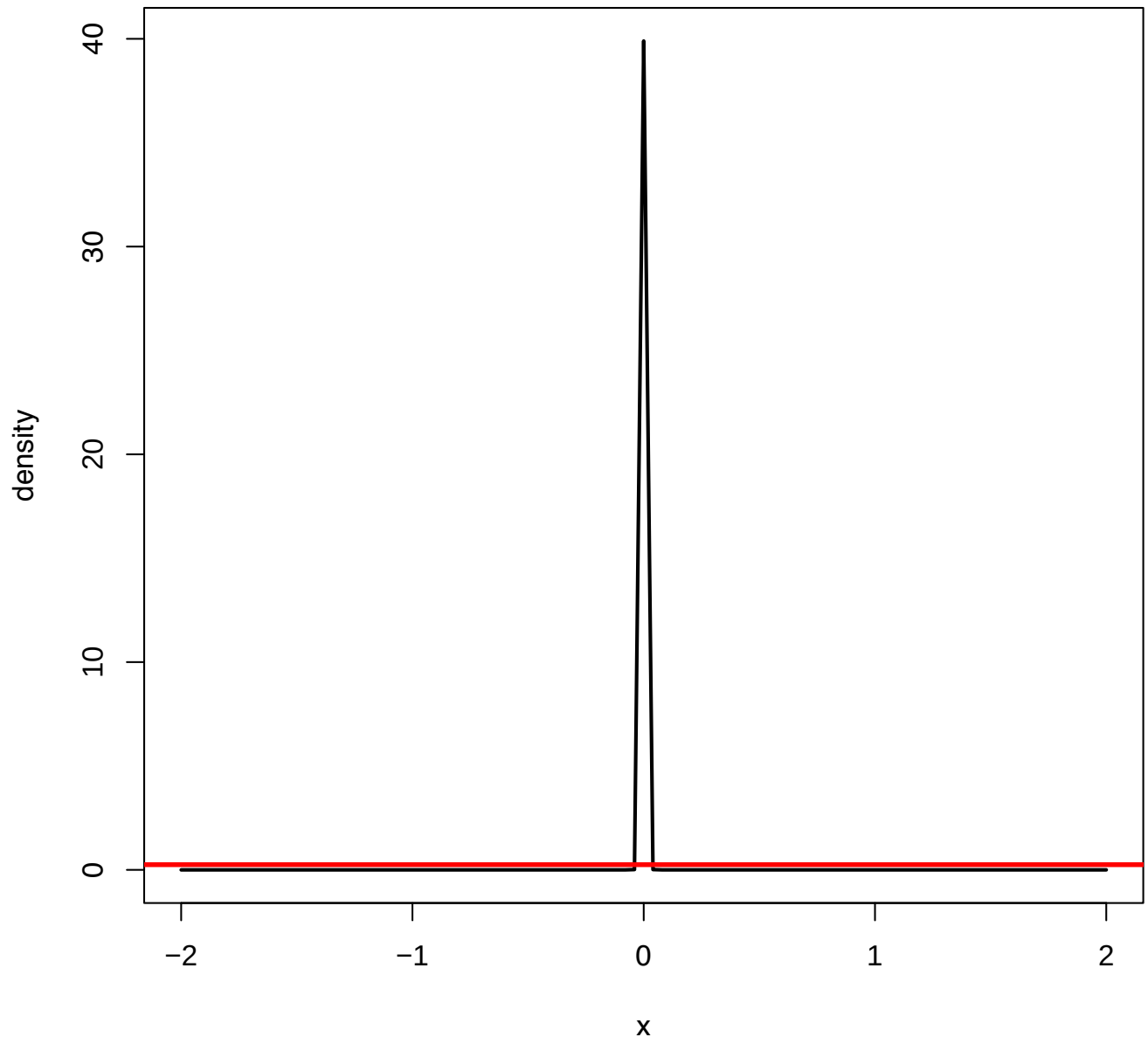
$$\frac{p(\theta^*|D,M)}{p(\theta \mid D,M)} = \frac{\mathbb{P}(D \mid \theta^*,M)p(\theta^* \mid M)}{\mathbb{P}(D \mid \theta,M)p(\theta \mid M)}$$

Bayesian model selection

Bayes Factor between two models:

$$B_{10} = \frac{\mathbb{P}(D \mid M_1)}{\mathbb{P}(D \mid M_0)}$$

We could estimate $\mathbb{P}(D|M_1)$ by drawing points from the prior on θ and calculating the mean likelihood.


Sharp posterior (black) and prior (red)
density
0
10
20
30
40
−2
−1
0
1
2
x

Drawing from the prior will often miss any of the trees with high posterior density – massive sample sizes would be needed.

Perhaps we can use samples from the posterior?

We can use the harmonic mean of the likelihoods of MCMC samples to estimate $\mathbb{P}(D|M_1)$.

However...

"The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever"

A post on Dr. Radford Neal's blog

`http://radfordneal.wordpress.com/2008/08/17/`

`the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever`

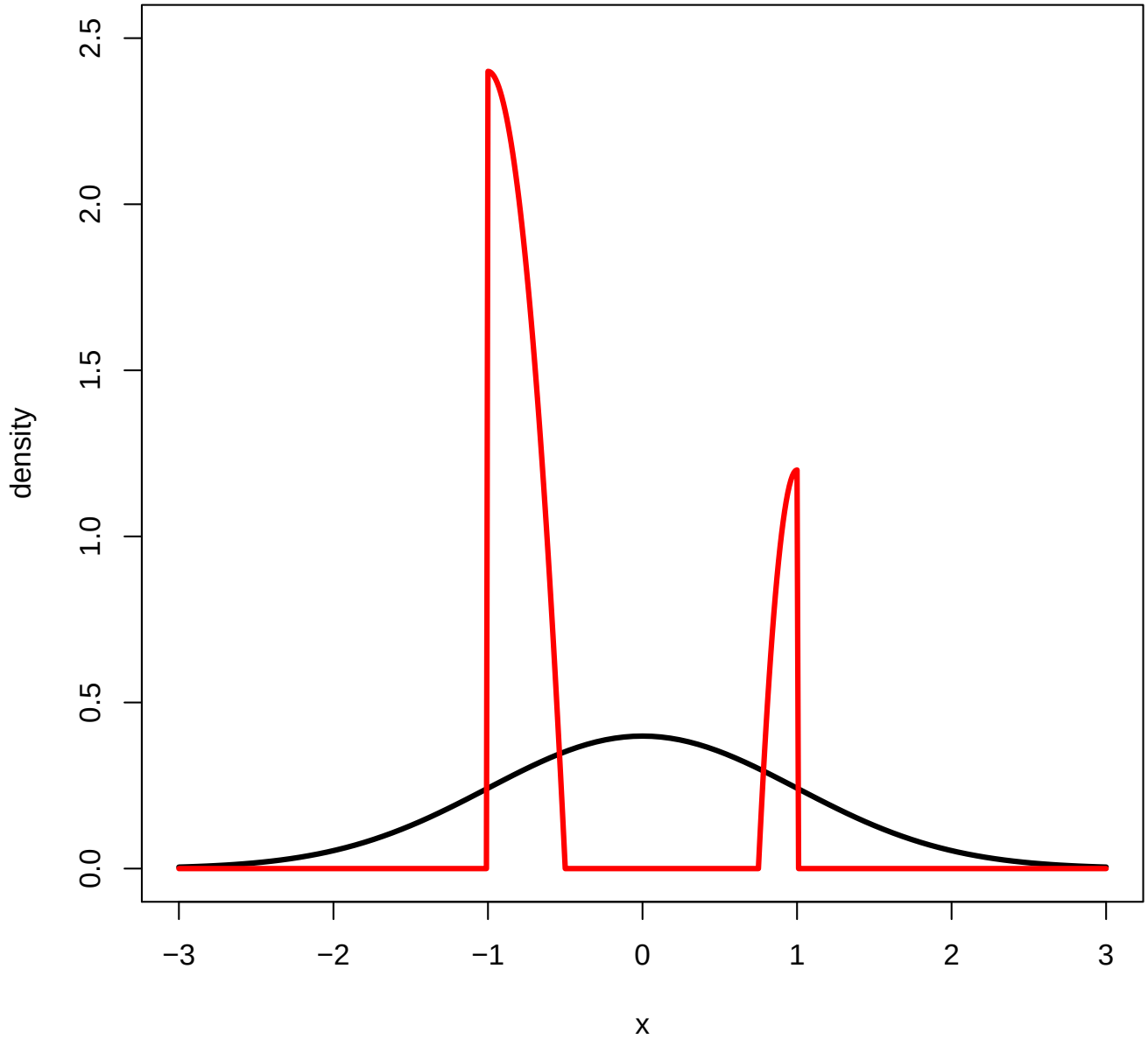
"The total unsuitability of the harmonic mean estimator should have been apparent within an hour of its discovery."

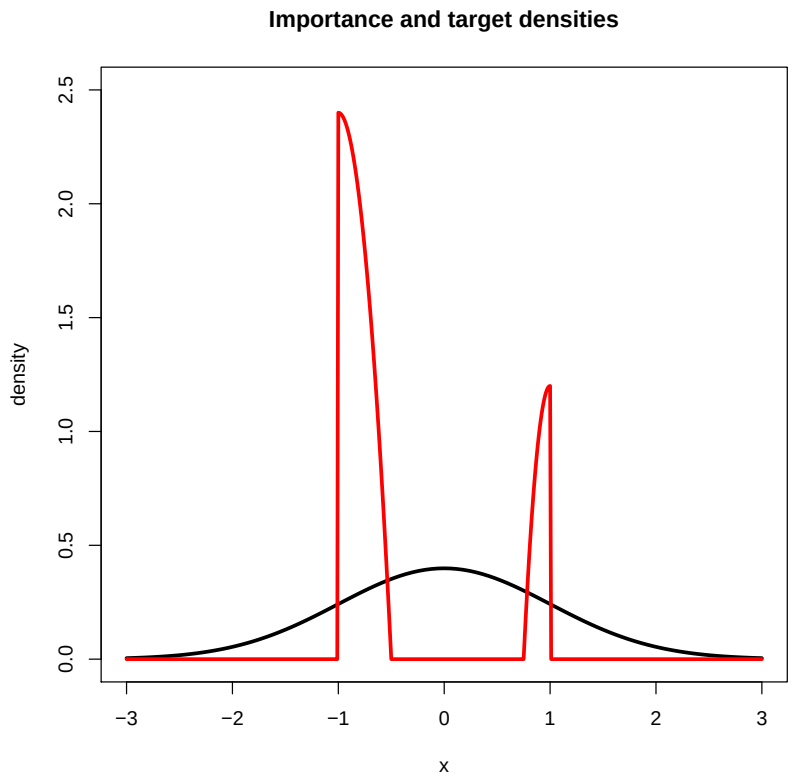
Harmonic mean estimator of the marginal likelihood

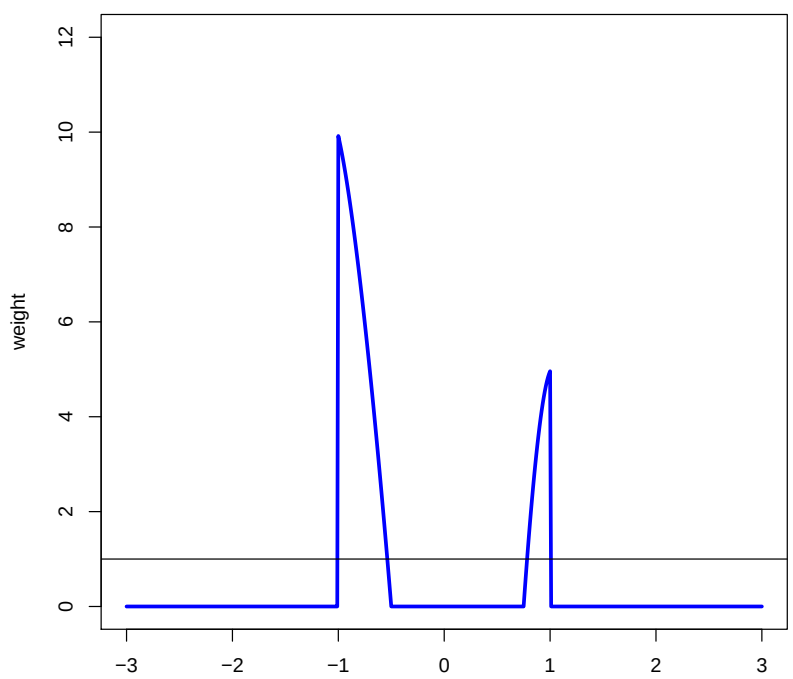
- appealing because it comes "for free" after we have sampled the posterior using MCMC,
- unfortunately the estimator can have a huge variance associated with it in some (very common) cases. For example if:
  - the vast majority of parameter space has very low likelihood, and
  - a very small region has high likelihoods.

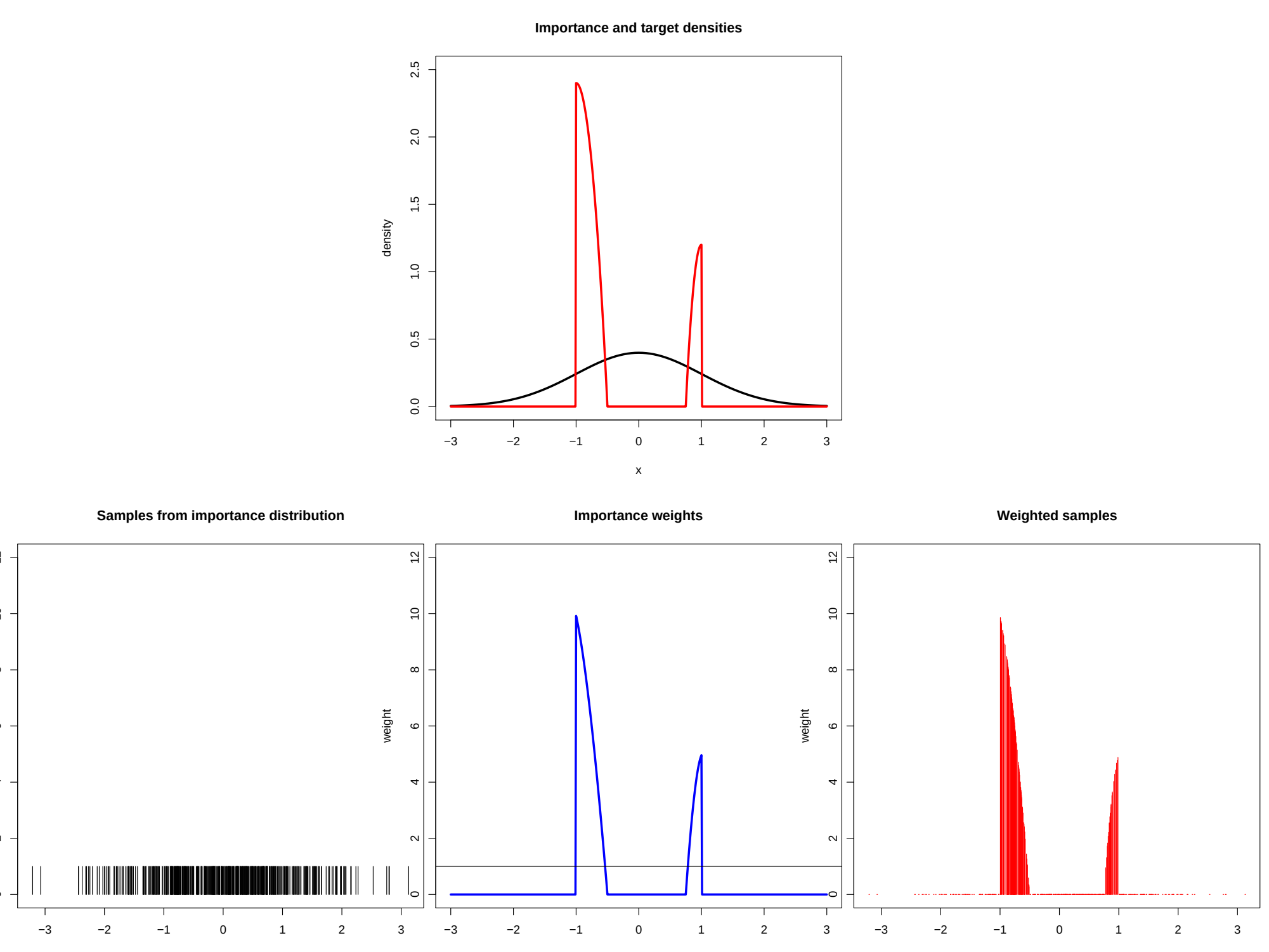
Importance sampling to approximate a difficult target distribution

1. Simulate points from an easy distribution.
2. Reweight the points by the ratio of densities between the easy and target distribution.
3. Treat the reweighted samples as draws from the target distribution.


density
0.0
0.5
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3
x


Importance and target densities
density
0.0
0.5
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3
x


Importance weights
weight
0
2
4
6
8
10
12
−3
−2
−1
0
1
2
3

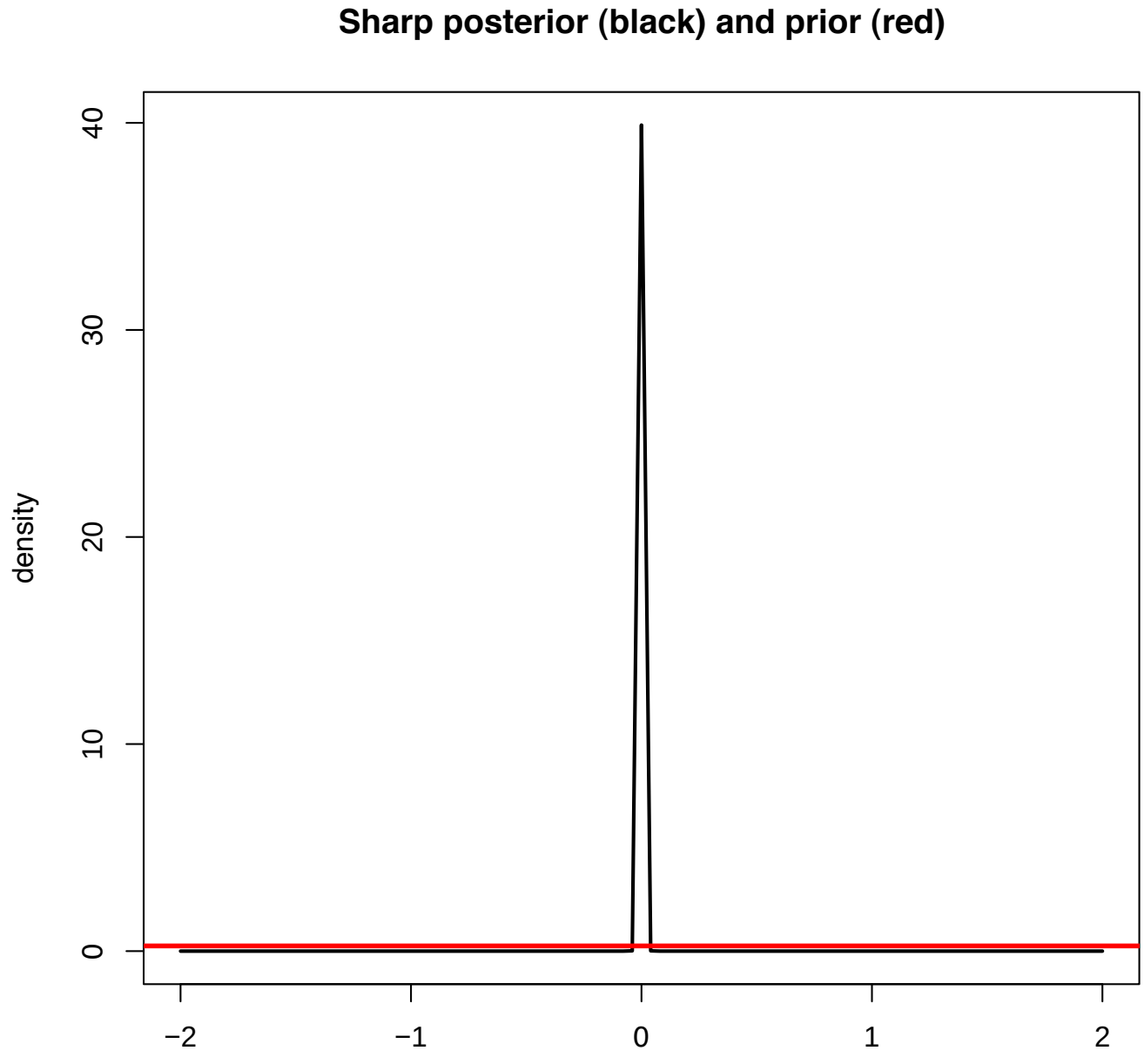

Importance and target densities
density
x
Samples from importance distribution
Importance weights
weight
Weighted samples
weight

Importance sampling

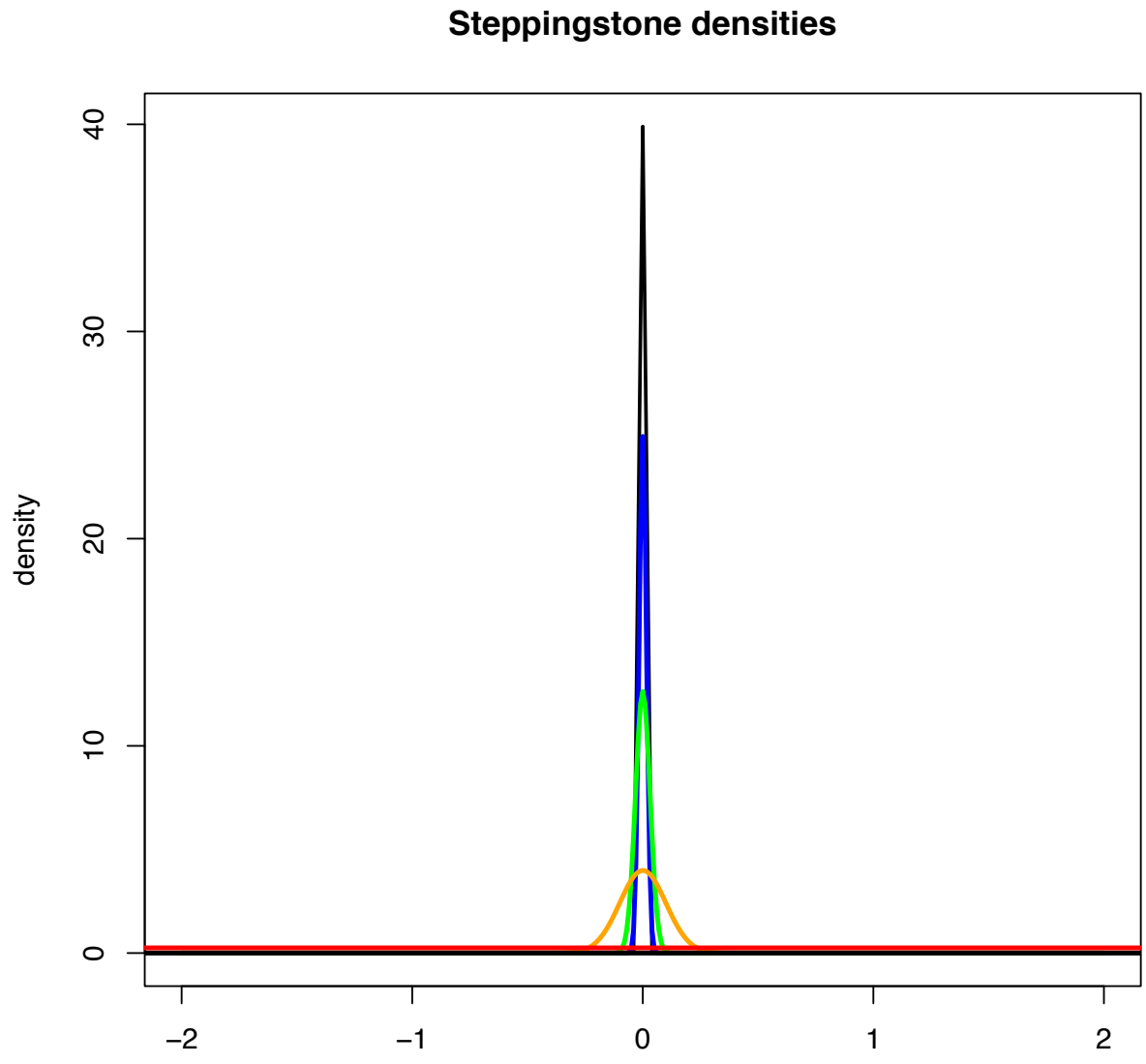
The method works well if the importance distribution is:

- fairly similar to the target distribution, and
- not "too tight" to allow sampling the full range of the target distribution

In phylogenetics our posterior distribution is too peaked and our prior is too vague to allow us to use them in importance sampling:



Steppingstone sampling uses a series of importance sampling runs



Steppingstone sampling (Xie *et al.* 2011, Fan *et al.* 2011) blends two distributions:

- the posterior, $\mathbb{P}(D \mid \theta, M_1)\mathbb{P}(\theta, M_1)$
- a tractable reference distribution, $\pi(\theta)$

$$p_\beta(\theta \mid D, M_1) \quad = \quad \frac{\left[\mathbb{P}(D \mid \theta, M_1)\mathbb{P}(\theta, M_1)\right]^\beta \left[\pi(\theta)\right]^{(1-\beta)}}{c_\beta}$$

$p_1(\theta \mid D, M_1)$ is the posterior.
c_1 is the marginal likelihood of the model.

$p_0(\theta \mid D, M_1)$ is the reference distribution.
c_0 is 1.

Steppingstone sampling (Xie *et al.* 2011, Fan *et al.* 2011) blends two distributions:

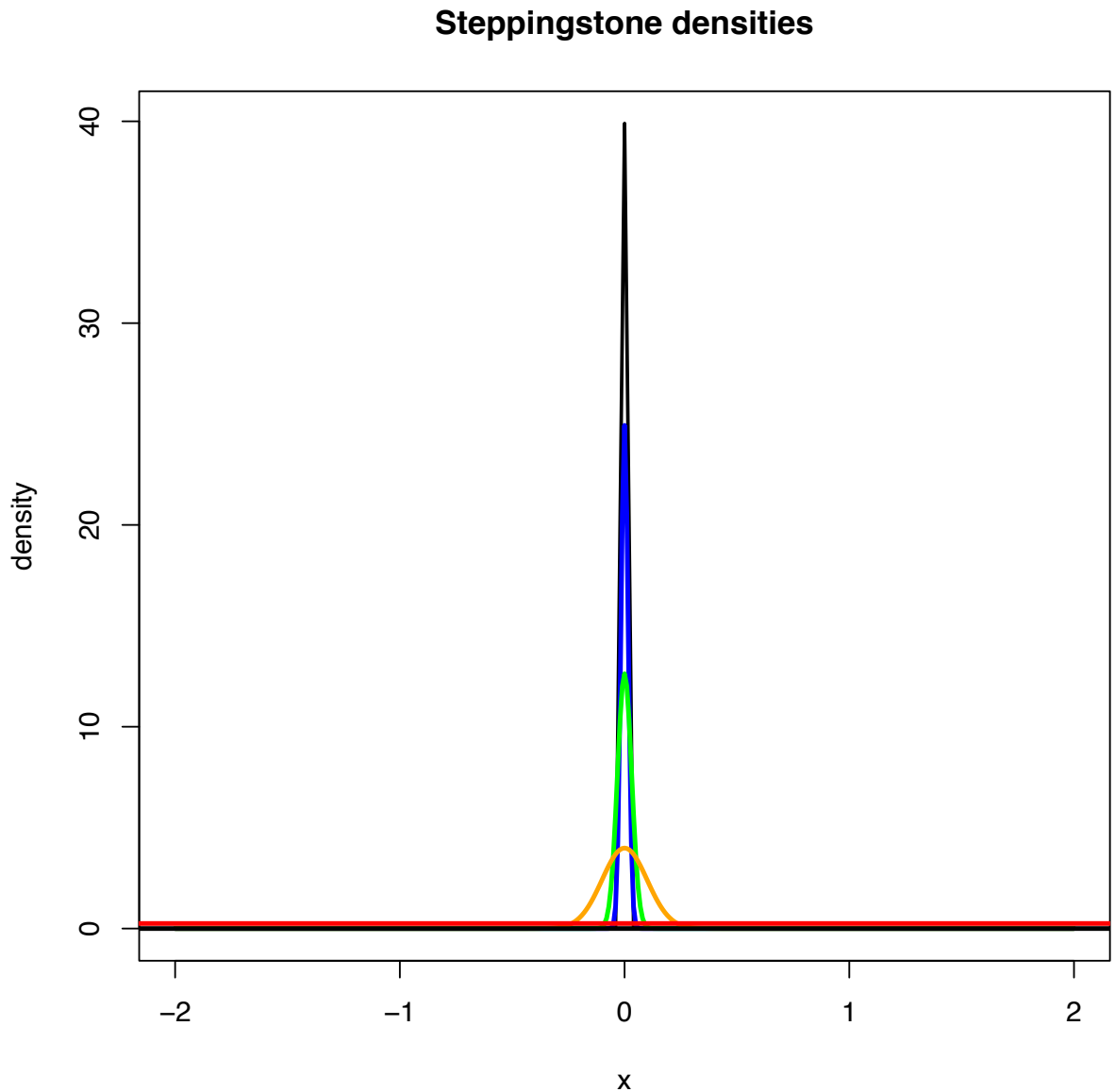
- the posterior, $\mathbb{P}(D \mid \theta, M_1)\mathbb{P}(\theta, M_1)$
- a tractable reference distribution, $\pi(\theta)$

$$p_\beta(\theta \mid D, M_1) = \frac{\left[\mathbb{P}(D \mid \theta, M_1)\mathbb{P}(\theta, M_1)\right]^\beta \left[\pi(\theta)\right]^{(1-\beta)}}{c_\beta}$$

$$\mathbb{P}(D \mid M_1) = \frac{c_1}{c_0} = \left(\frac{c_1}{c_{0.38}}\right)\left(\frac{c_{0.38}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.01}}\right)\left(\frac{c_{0.01}}{c_0}\right)$$

$$= \left(\frac{c_1}{\cancel{c_{0.38}}}\right)\left(\frac{\cancel{c_{0.38}}}{\cancel{c_{0.1}}}\right)\left(\frac{\cancel{c_{0.1}}}{\cancel{c_{0.01}}}\right)\left(\frac{\cancel{c_{0.01}}}{c_0}\right)$$

Run MCMC with different β values



$$\mathbb{P}(D \mid M) = \left(\frac{\mathbb{P}(D|M)}{c_{0.38}}\right)\left(\frac{c_{0.38}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.01}}\right)\left(\frac{c_{0.01}}{1}\right)$$

Photo by Johan Nobel http://www.flickr.com/photos/43147325@N08/4326713557/ downloaded from Wikimedia

Reference distributions in Steppingstone sampling

In the original steppingstone (Xie et al 2011):

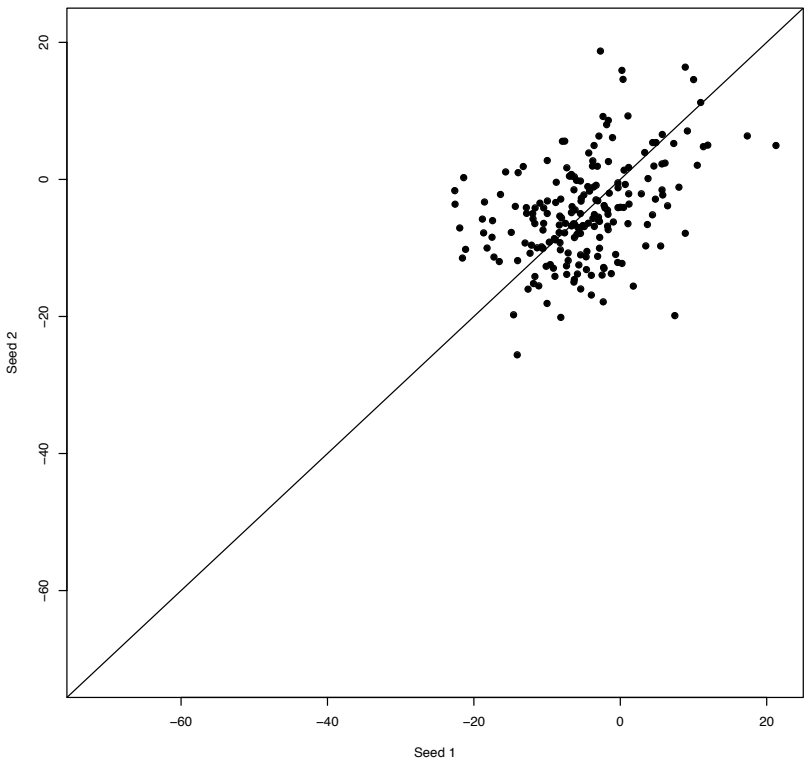
$$\text{reference} = \text{prior}$$

In generalized steppingstone (Fan et al 2011) it can be any distribution:

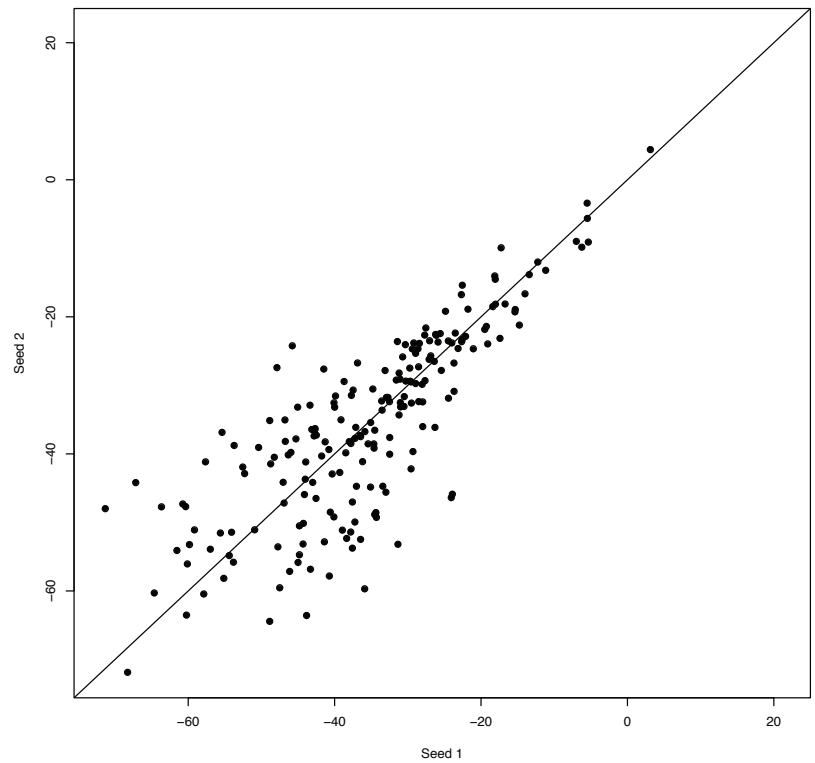
- Should be centered around the areas with high probability,
- Must be a probability distribution with a known normalizing constant,
- Should be easy to draw sample from.

The log Bayes Factor for a complex model compared to a simple (true) model. Estimated twice (by Fan et al, 2011)

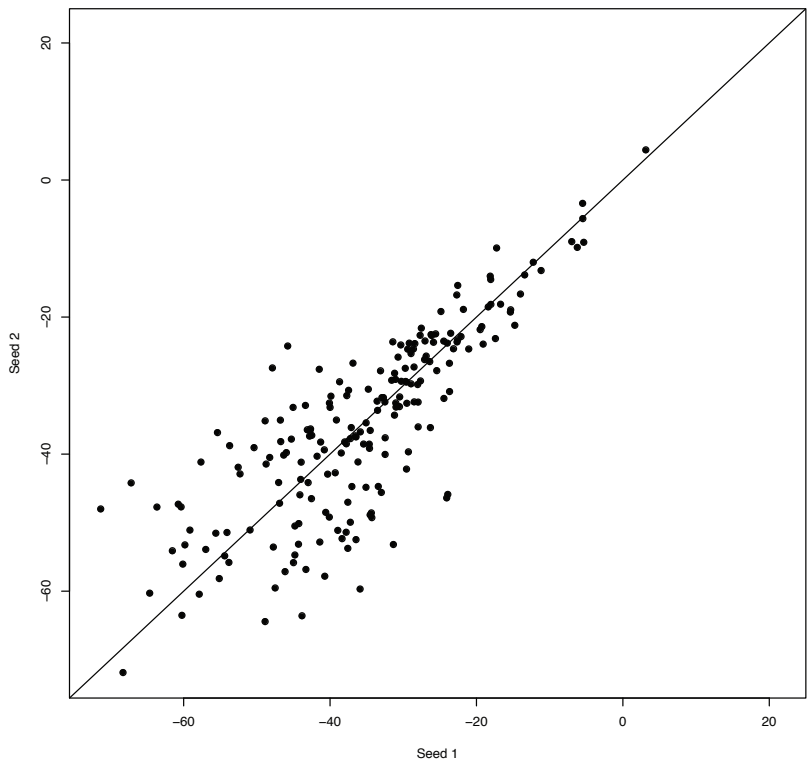
Harmonic mean:



Original Steppingstone:



Original Steppingstone:



Generalized Steppingstone:

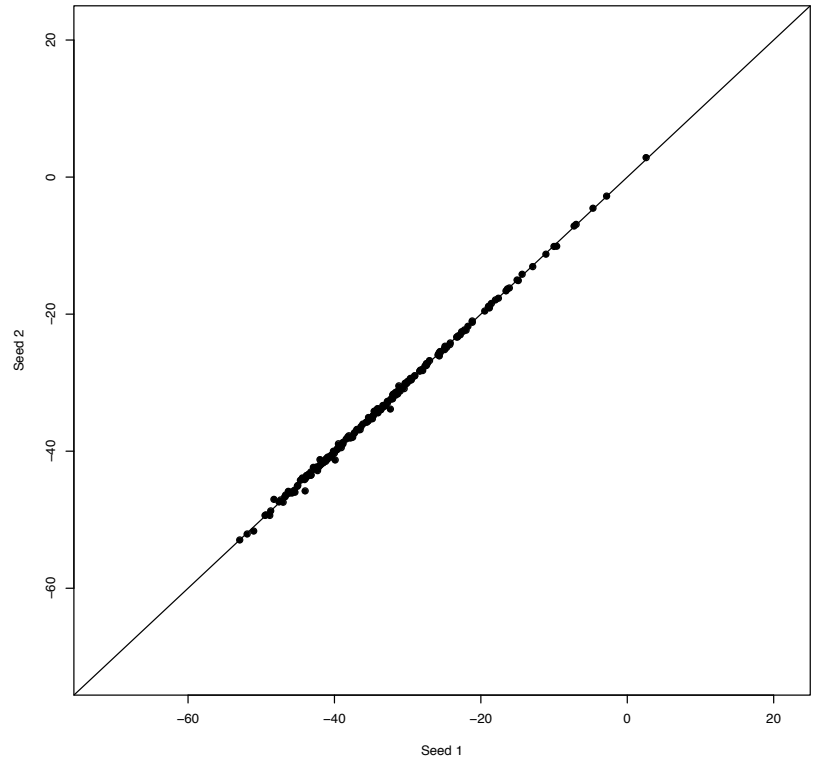


Figure from Fan et al., 2011

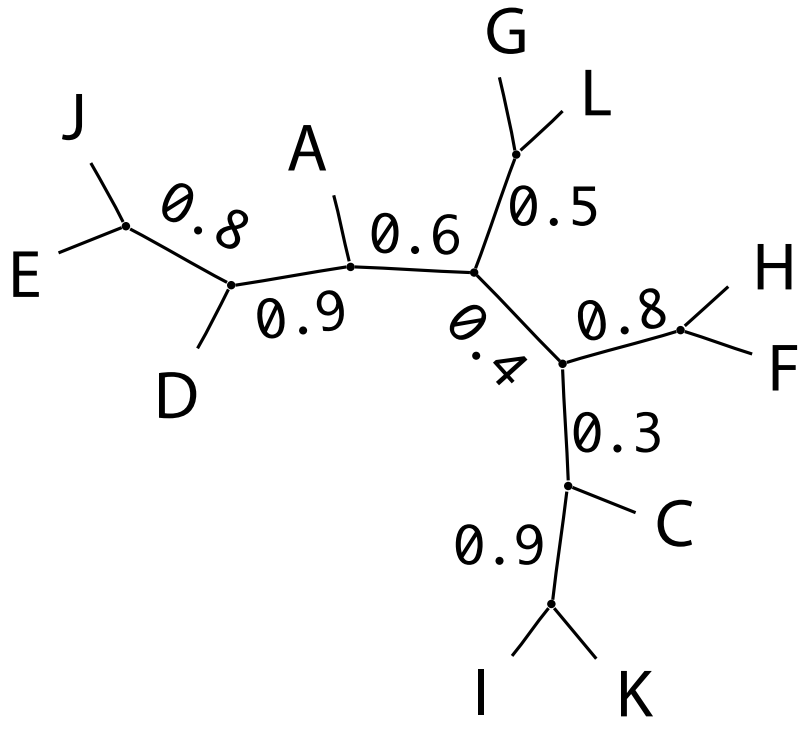
Steppingstone sampling when the tree is not known

- generalized steppingstone assumes a fixed tree,
- the original steppingstone can very slow when the tree is not known.

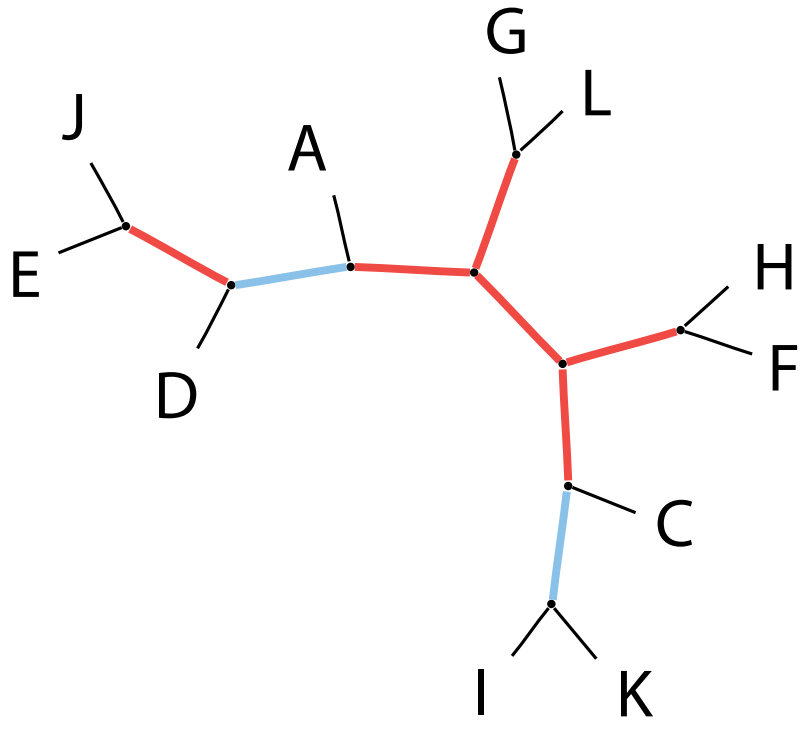
Tree-Centered Independent-Split-Probability (TCISP) distribution

Input: a tree with probabilities for each split (from a pilot MCMC run).

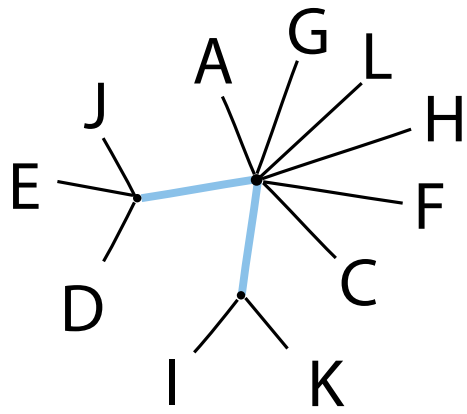
Output: a probability distribution over all tree topologies.

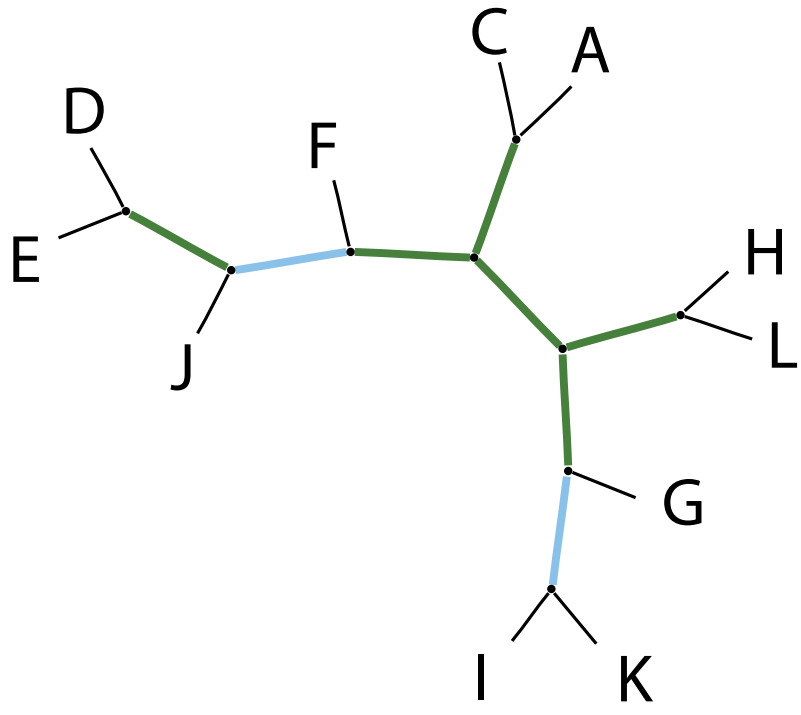


Input: a focal tree
with split probabilities
to center the distribution

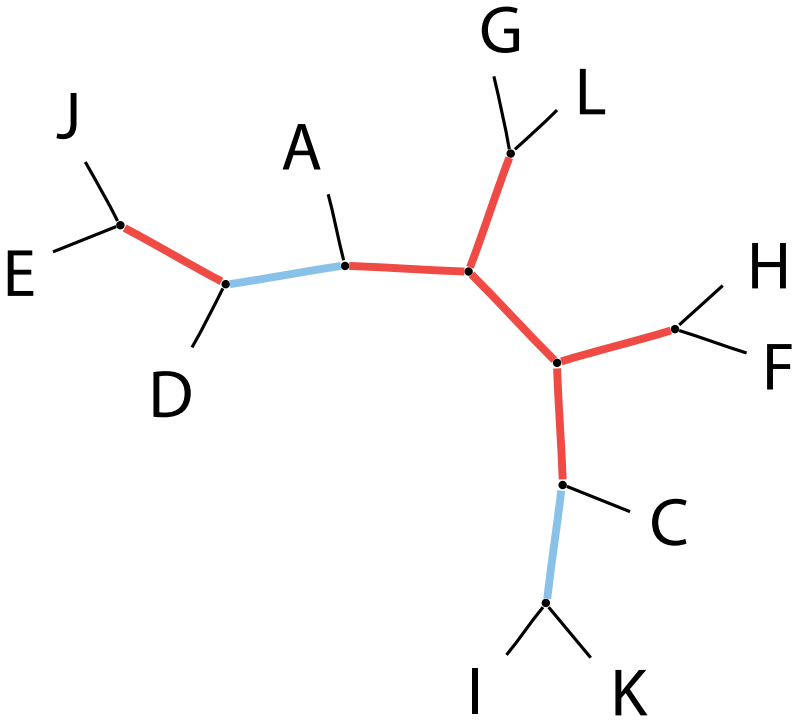


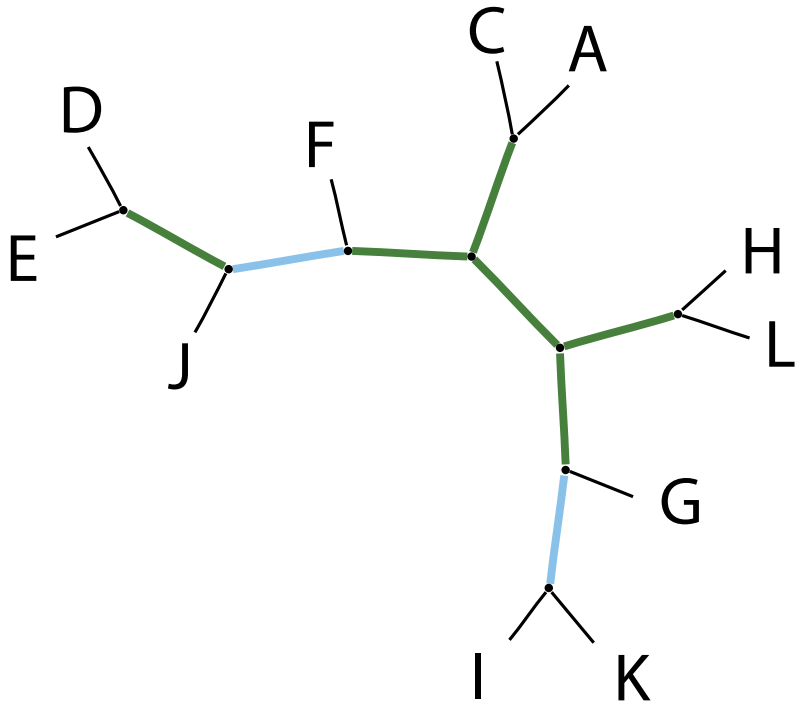
The tree we will draw
will display the blue splits
and will not display the red splits


G
A
L
J
H
E
F
C
D
I
K

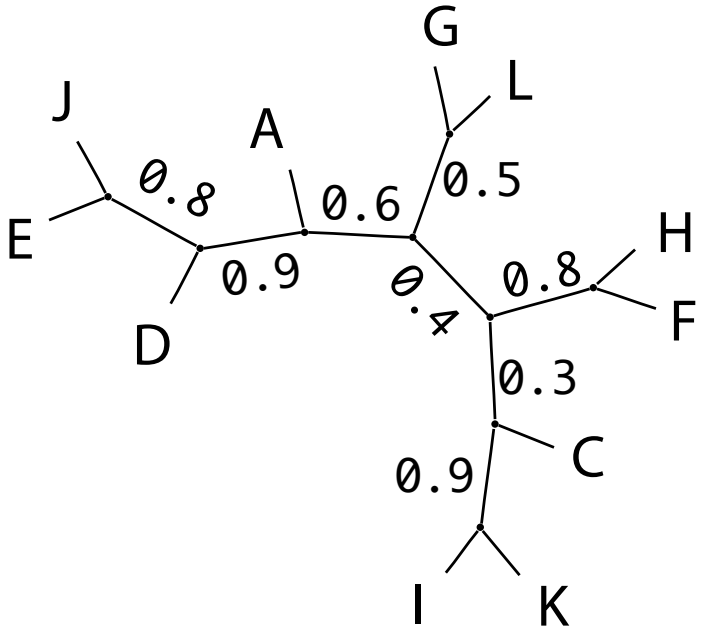


One of the many resolutions
which avoid the red splits

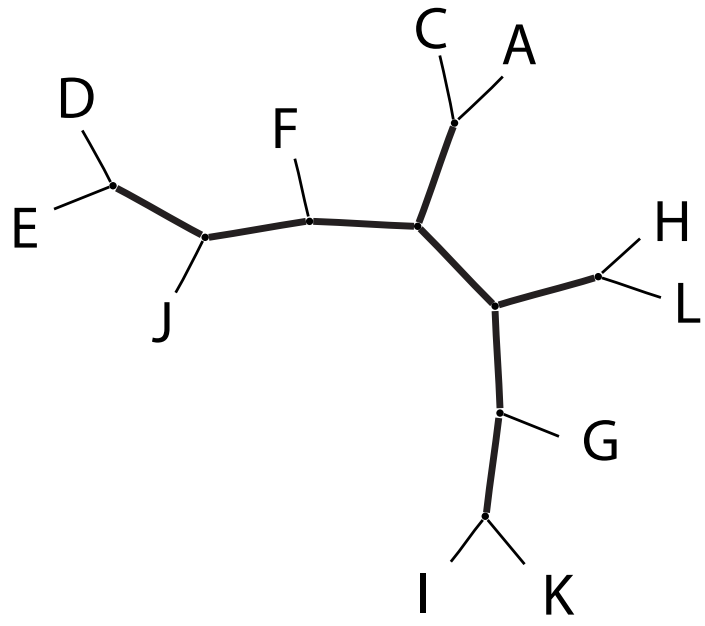

G
L
J
A
E
H
F
D
C
I
K


C
A
D
F
E
H
L
J
G
I
K

Calculating a tree's probability

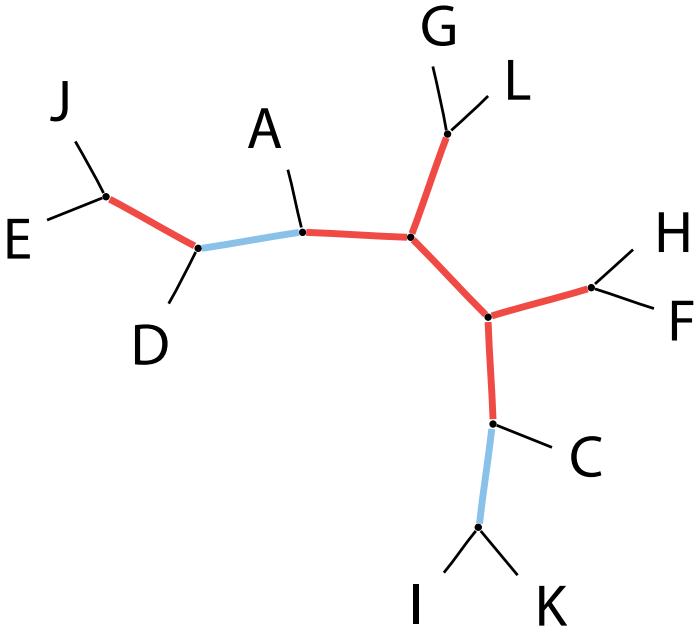


focal tree

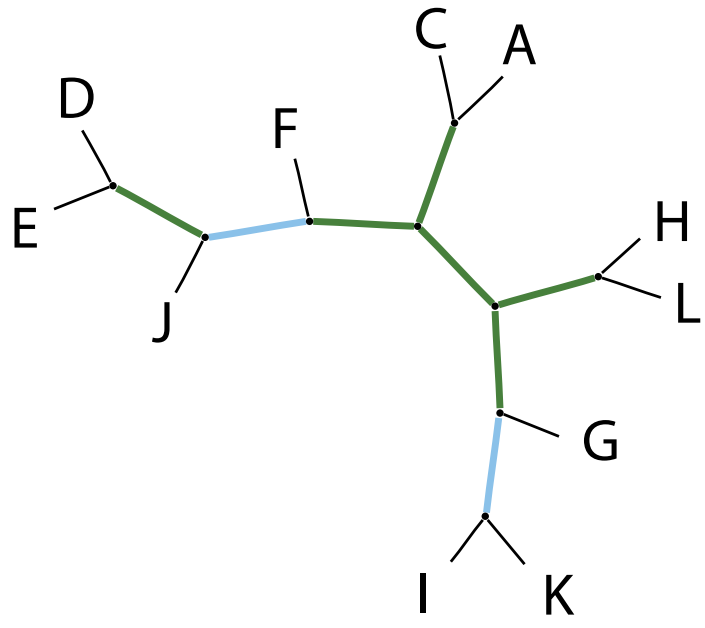


tree to score

Calculating a tree's probability

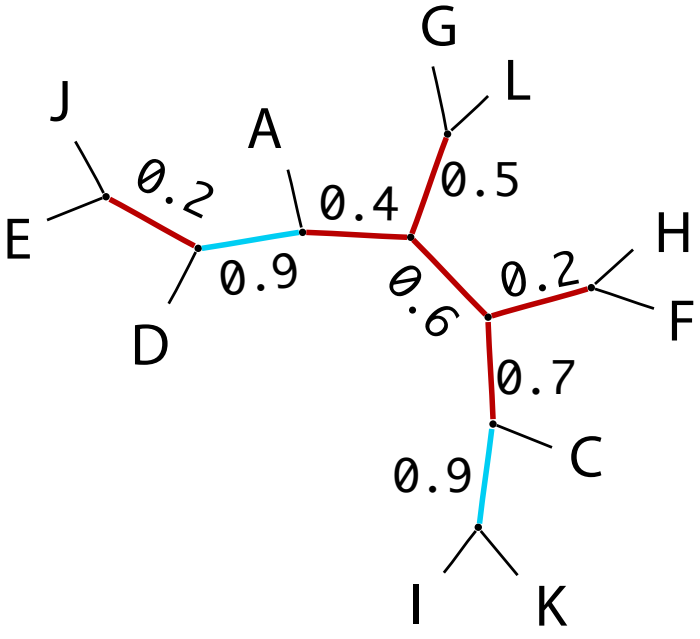


focal tree

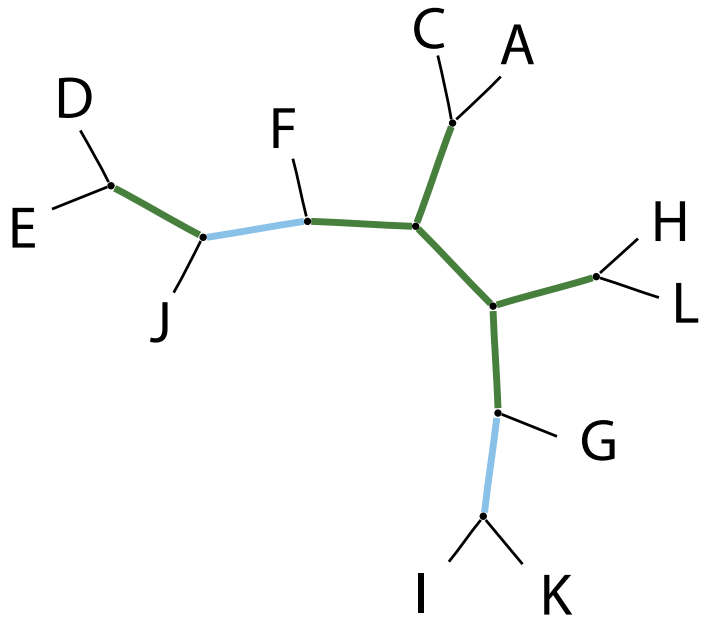


tree to score

Calculating a tree's probability

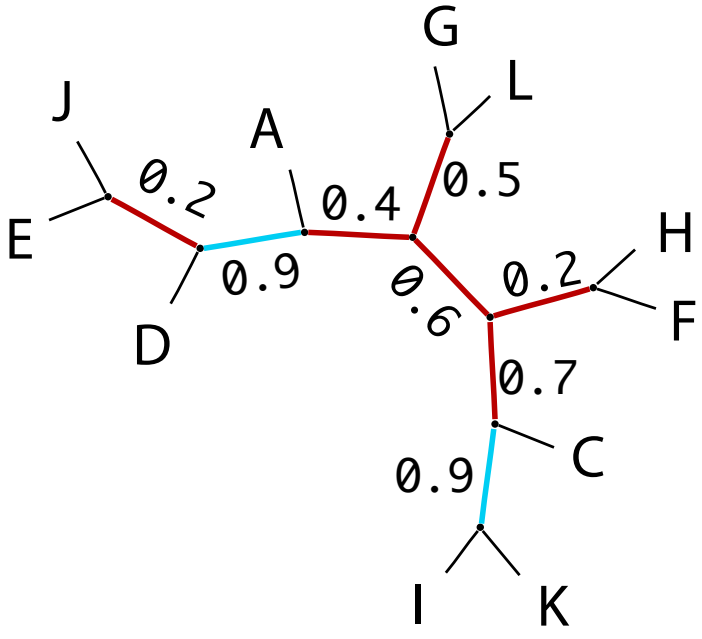


focal tree

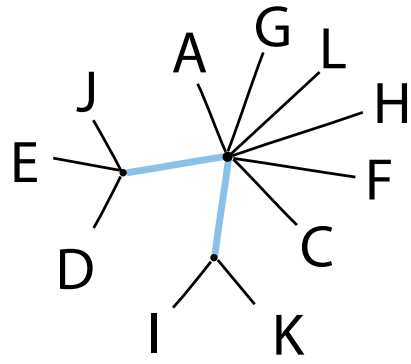


tree to score

Calculating a tree's probability



focal tree



tree with shared splits

$$\mathbb{P} = 0.2 \times 0.9 \times 0.4 \times 0.5 \times 0.6 \times 0.2 \times 0.7 \times 0.9$$

Counting trees the max. distance from a tree

- Bryant and Steel(2009) provide an $\mathcal{O}(n^5)$ algorithm.
- Bryant contributed an $\mathcal{O}(n^2)$ algorithm.

Multitree steppingstone sampling

- Has been validated on small data sets.
- Minimum running time is unknown.
- Appears to be a practical way to approximate a model's marginal likelihood.
- The TCISP distribution could be useful in other context.
- The method assumes unrooted, fully resolved trees,
- implemented in phycas (`http://www.phycas.org`)

Holder, Lewis, Swofford, Bryant (*in prep*)

Marginal likelihoods

- **Harmonic mean estimator is not reliable**
- AIC is preferable to the harmonic mean estimator (Guy Baele *et al* MBE 2012)
- Thermodynamic Integration (aka Path Sampling) is an accurate method (Lartillot and Philippe, 2006; Rodrigue and Aris-Brousou)
- Model-jumping approaches avoid the need to calculate marginal likelihoods.
- Steppingstone and/or Path Sampling are available in MrBayes, Beast, *Beast, Migrate, and other Bayesian software for evolutionary analyses

Thanks!

Thanks to the organizers, NSF, and to you (for listening to stats on a Saturday morning)

This slide shows the math to demonstrate that we can use the harmonic mean to estimate the marginal likelihood, $P(D)$. Consider, a model that has a discrete parameter v that can take two values (r and b). h in the harmonic mean of the likelihoods for parameters values sampled from the posterior distribution

$$
\begin{aligned}
h &= \frac{1}{\mathbb{P}(v=r|D)\left[\frac{1}{\mathbb{P}(D|v=r)}\right] + \mathbb{P}(v=b|D)\left[\frac{1}{\mathbb{P}(D|v=b)}\right]} \\
&= \frac{1}{\left[\frac{\mathbb{P}(v=r)\mathbb{P}(D|v=r)}{\mathbb{P}(D)}\right]\left[\frac{1}{\mathbb{P}(D|v=r)}\right] + \left[\frac{\mathbb{P}(v=b)\mathbb{P}(D|v=b)}{\mathbb{P}(D)}\right]\left[\frac{1}{\mathbb{P}(D|v=b)}\right]} \\
&= \frac{1}{\left[\frac{\mathbb{P}(v=r)}{\mathbb{P}(D)}\right] + \left[\frac{\mathbb{P}(v=b)}{\mathbb{P}(D)}\right]} \\
&= \frac{\mathbb{P}(D)}{\mathbb{P}(v=r) + \mathbb{P}(v=b)} \\
&= \mathbb{P}(D)
\end{aligned}
$$

The harmonic mean as an importance sampling estimator

The marginal likelihood is the expected value of the likelihood over points drawn from the prior:

$$\mathbb{P}(D) = \mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] \quad = \quad \int \mathbb{P}(D|\theta)\mathbb{P}(\theta)d\theta$$

In importance sampling, we draw parameter values $\boldsymbol{\theta}$ from a different density, $g(\theta)$, but multiple the points by a weight, w:

$$\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] \quad = \quad \frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(D|\theta_i)w_i}{b}$$

where b is a normalization constant.

The appropriate weight is the ratio of the target density and our importance density:

$$w_i = \frac{\mathbb{P}(\theta_i)}{g(\theta_i)}$$

And it turns out that the normalization constant is simply a sum of the importance weights:

$$b = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{g(\theta_i)}$$

$$\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}{g(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{g(\theta_i)}}$$

$$\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}{g(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{g(\theta_i)}}$$

If the importance distribution, $g(\theta)$, is the prior then the importance weights are all 1:

$$\begin{aligned}\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] &= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}{\mathbb{P}(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{\mathbb{P}(\theta_i)}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(D|\theta_i)\end{aligned}$$

Recall that the points, $\boldsymbol{\theta}$, are draw by sampling from the importance distribution, so this sum of likelihoods is already weighted by the prior.

$$\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}{g(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{g(\theta_i)}}$$

If the importance distribution, $g(\theta)$ is the posterior then:

$$\begin{aligned}\mathbb{E}_{p(\theta)}[\mathbb{P}(D|\theta)] &= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}} \\ &= \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\theta_i)}{\mathbb{P}(D|\theta_i)\mathbb{P}(\theta_i)}} \\ &= \frac{n}{\sum_{i=1}^{n}\frac{1}{\mathbb{P}(D|\theta_i)}}\end{aligned}$$

This is the justification of the harmonic mean estimator of the marginal likelihood.

Lartillot and Philippe's thermodynamic integration

Like the steppingstone sampler, the thermodyanmic integration method (or path sampling) use power posterior densities with $0 \leq \beta \leq 1$:

$$p_\beta(\theta|D) = \frac{\mathbb{P}(D|\theta)^\beta \mathbb{P}(\theta)}{c_\beta}$$

with $c_0 = 0$ and $c_1 = \mathbb{P}(D)$.
Lartillot and Philippe showed that

$$\frac{\partial \ln c_\beta}{\partial \beta} = \mathbb{E}_{p_\beta}\left[\frac{\partial \ln \left[\mathbb{P}(D|\theta)^\beta \mathbb{P}(\theta)\right]}{\partial \beta}\right]$$

Note that by definition of a definite integral:

$$\int_0^1 \frac{\partial \ln c_\beta}{\partial \beta} d\beta = \ln c_1 - \ln c_0 = \ln \mathbb{P}(D)$$

Thus,

$$\ln \mathbb{P}(D) = \int_0^1 \mathbb{E}_{p_\beta} \left[\frac{\partial \ln \left[\mathbb{P}(D|\theta)^\beta \mathbb{P}(\theta) \right]}{\partial \beta} \right] d\beta$$

By differentiating we see that:

$$\begin{aligned} \frac{\partial \ln \left[\mathbb{P}(D|\theta)^\beta \mathbb{P}(\theta) \right]}{\partial \beta} &= \frac{\partial [\ln \mathbb{P}(\theta) + \beta \ln \mathbb{P}(D|\theta)]}{\partial \beta} \\ &= \ln \mathbb{P}(D|\theta) \end{aligned}$$

The integration is not analytically tractable, but we can calculate $\mathbb{E}_{p_\beta} [\ln \mathbb{P}(D|\theta)]$ by conducting MCMC at the a power posterior distribution p_β and taking an average of the log-likelihoods.
Then we can use standard numerical integration techniques to estimate the integral.

This integration is not analytically tractable, but we can calculate $\mathbb{E}_{p_\beta}\left[\ln \mathbb{P}(D|\theta)\right]$ by conducting MCMC at the a power posterior distribution p_β and taking an average of the log-likelihoods.
Then we can use standard numerical integration techniques to estimate the integral.