Phycas demonstration – polytomy priors, slice sampling, and steppingstone sampling

Mark Holder
Department of Ecology and Evolutionary Biology
University of Kansas
Lawrence, Kansas

marginal likelihood estimation

In ML model selection we judge models by their ML score and the number of parameters. In Bayesian context we:

- Use model averaging if we can "jump" between models (reversible jump methods, Dirichlet Process Prior, Bayesian Stochastic Search Variable Selection),
- Compare models on the basis of their marginal likelihood.

The Bayes Factor between two models:

$$B_{10} = \frac{p(D|M_1)}{p(D|M_0)}$$

is a form of likelihood ratio.

Bayes factor:

$$B_{10} = \frac{p(D|M_1)}{p(D|M_0)}$$

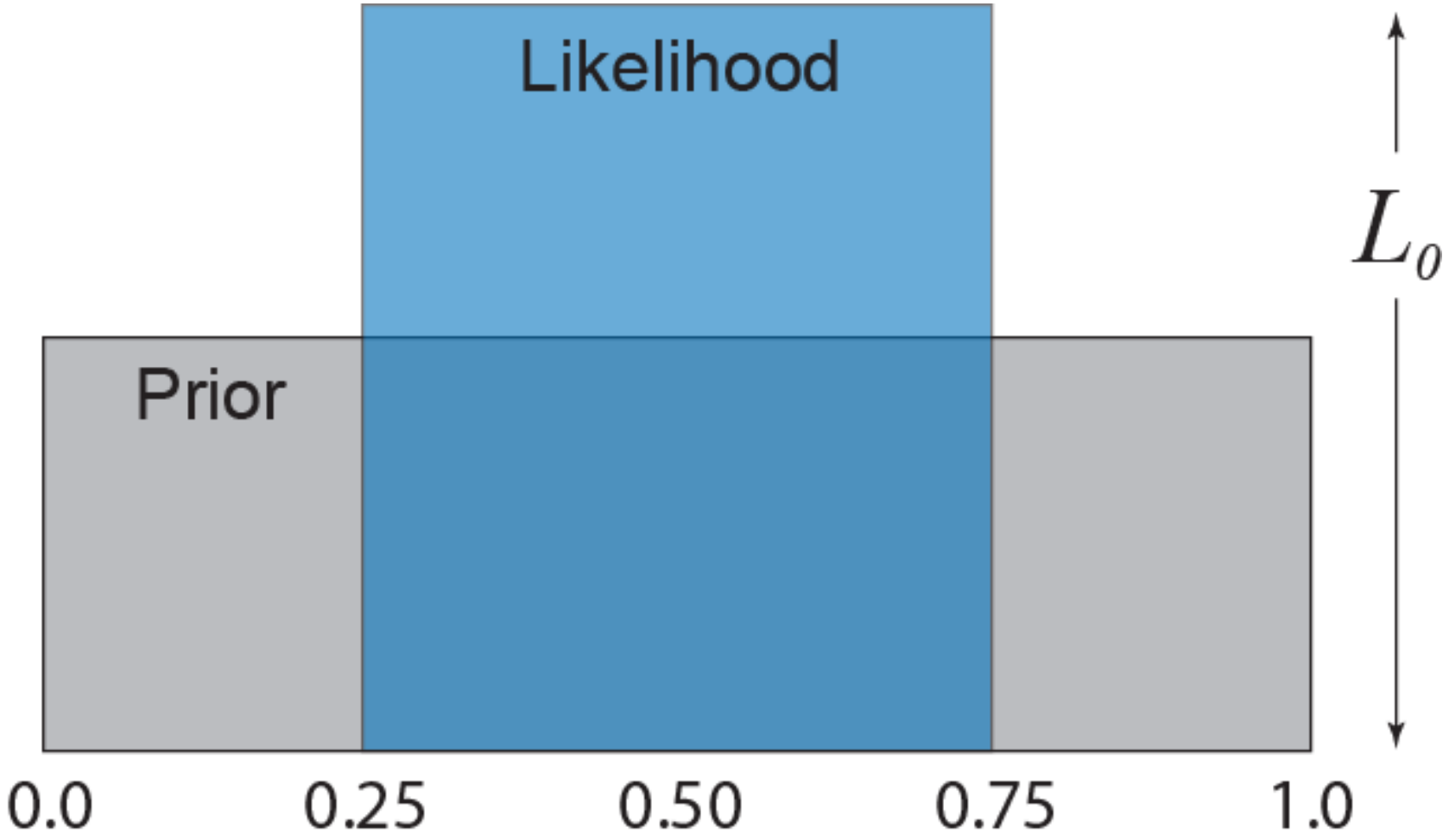
Let's unpack, $p(D|M_1)$, the marginal probability of the data under the model, M_1:

$$p(D|M_1) = \int p(D|\theta, M_1)p(\theta)d\theta$$

where θ is the set of parameters in the model.

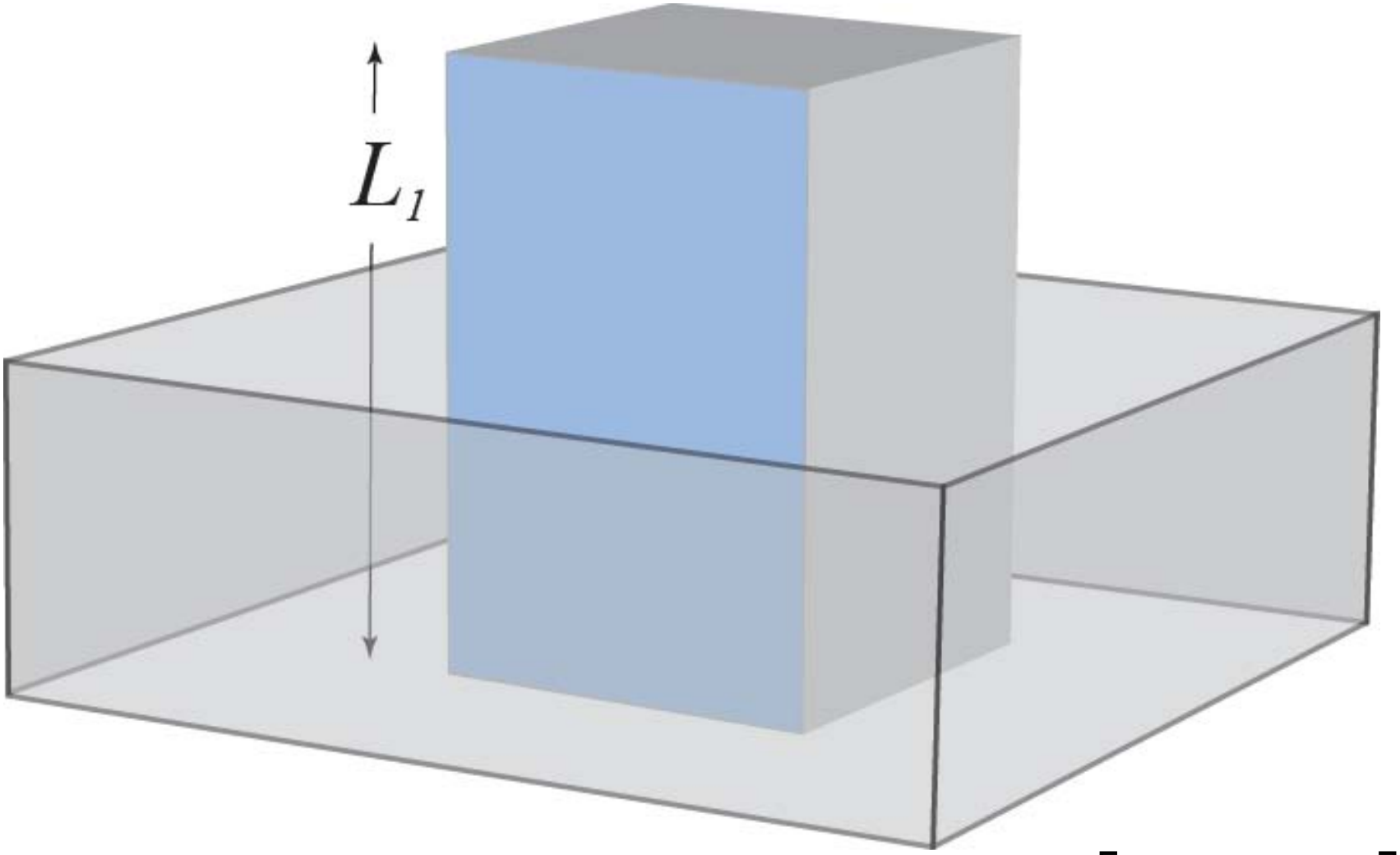
(The next slides are from Paul Lewis)

Marginal likelihood (1-param. model)



Average likelihood = $\left(\frac{1}{2}\right) L_0 + \left(\frac{1}{2}\right) (0)$

Marginal likelihood (2-param. model)

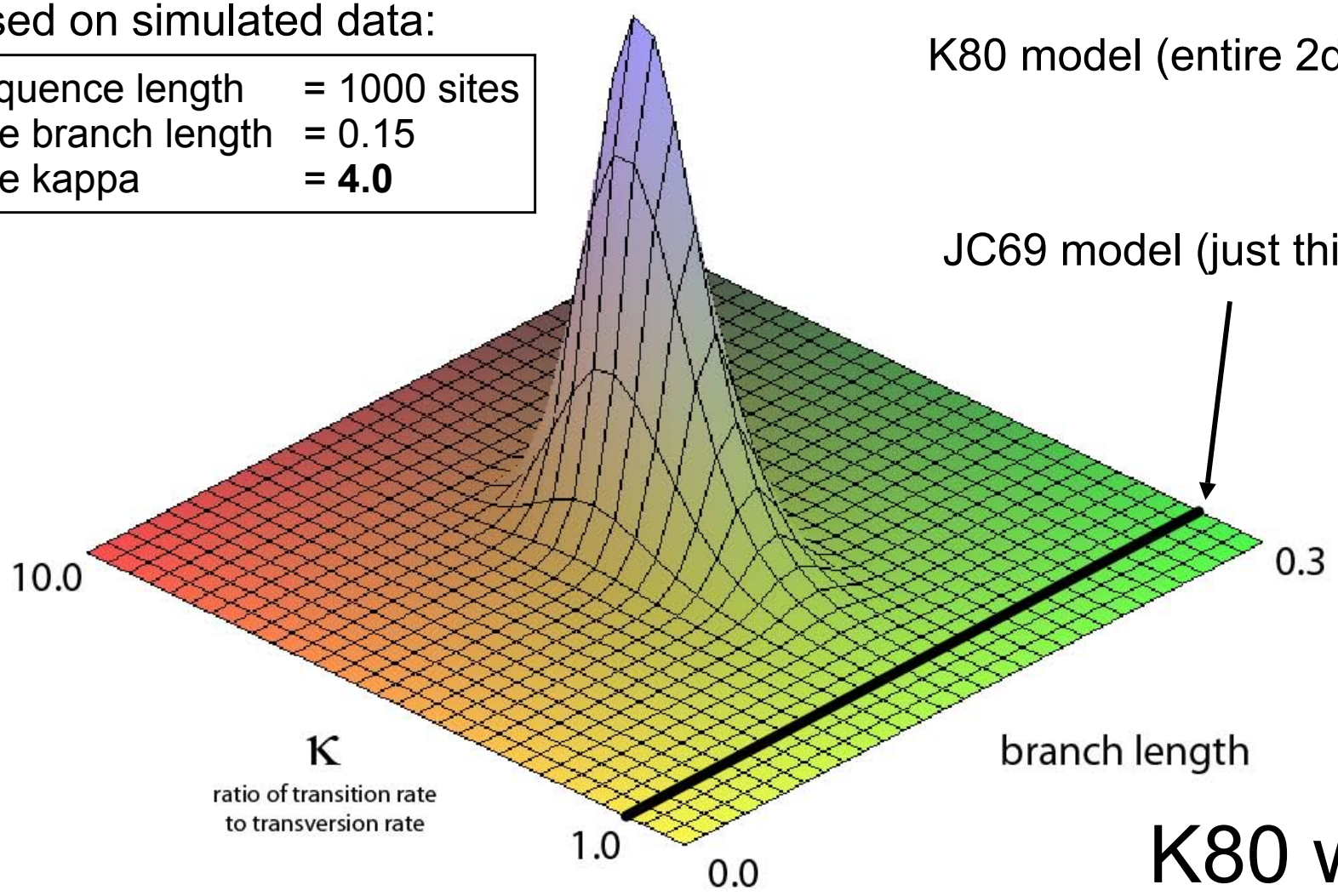


Average likelihood = $\left(\frac{1}{2}\right)^2 L_1 + \left[1 - \left(\frac{1}{2}\right)^2\right](0)$

Likelihood Surface when K80 true

Based on simulated data:

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **4.0**

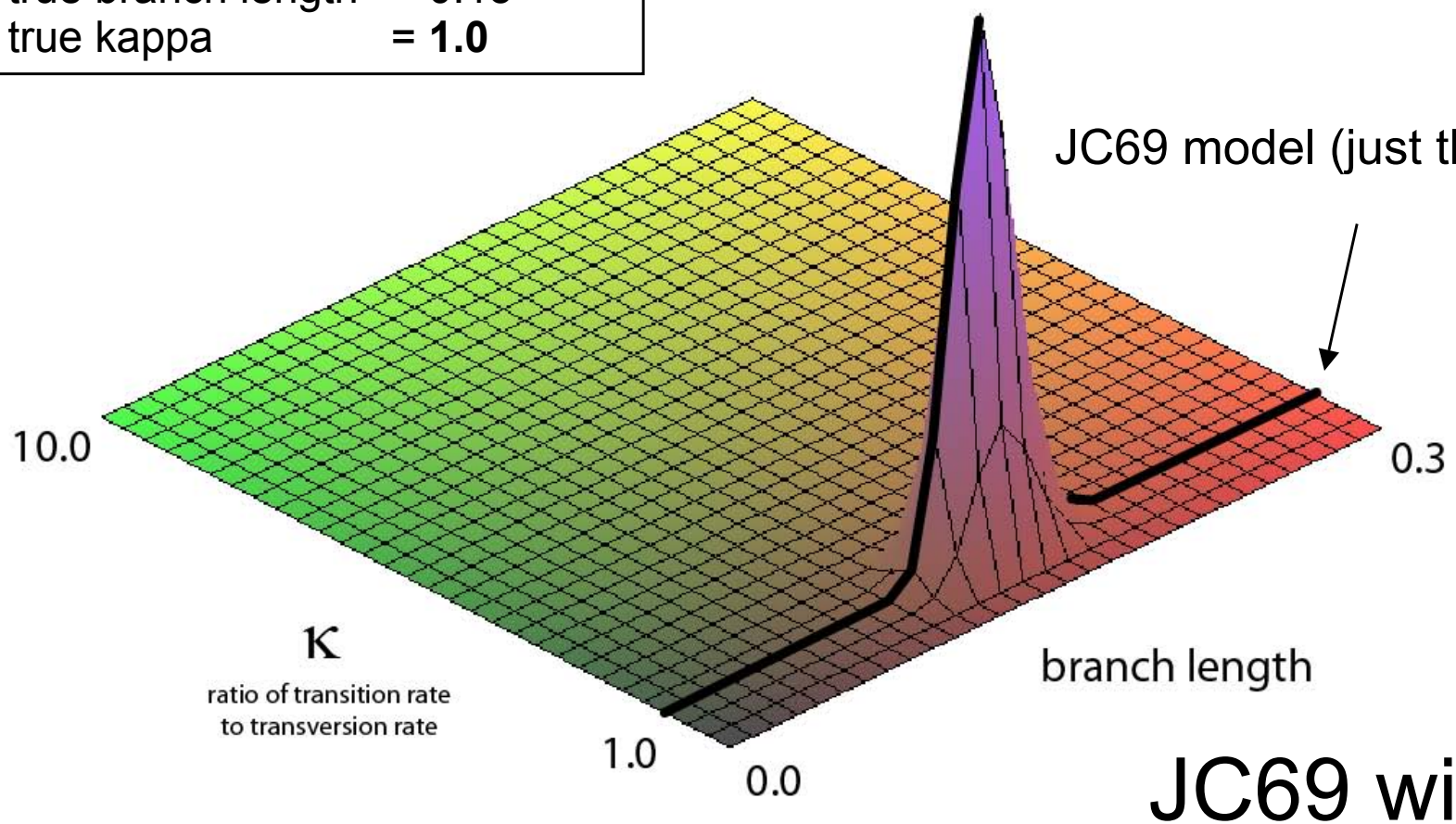


K80 wins

Likelihood Surface when JC true

Based on simulated data:

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **1.0**



JC69 wins

Important point: (but not the focus of the lab) Bayes Factor comparison remove the effect of the prior on the model itself, but the priors on nuisance parameters still matter!

Think about your priors - using a very parameter-rich model may not be overparameterized if you have prior knowledge about the parameter values.

Marginal likelihood formula (again)

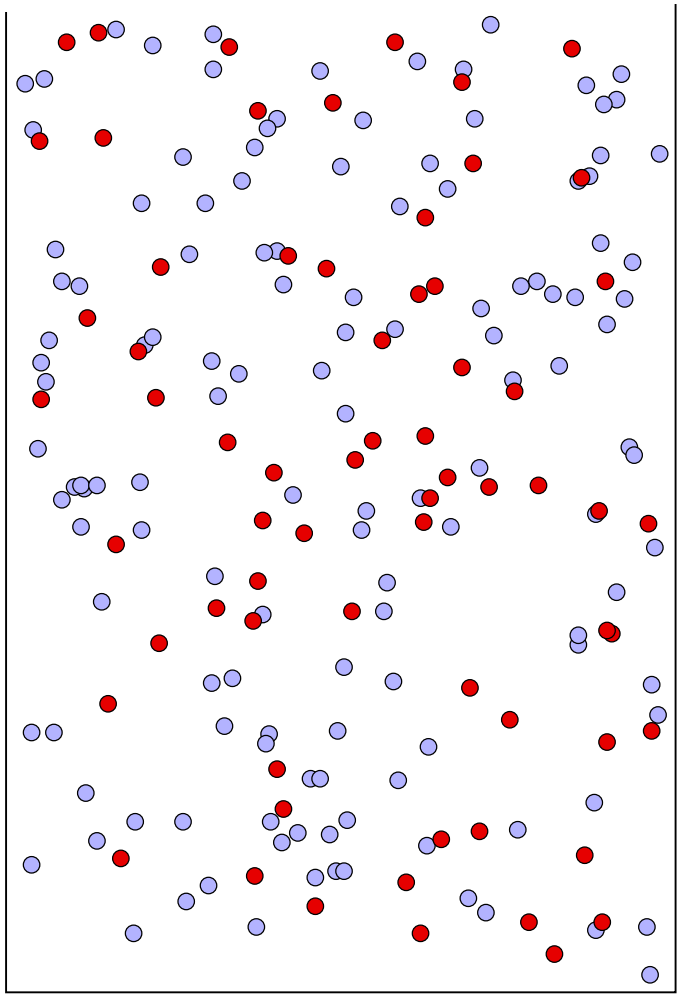
$$p(D|M_1) = \int p(D|\theta, M_1)p(\theta)d\theta$$

where θ is the set of parameters in the model.

We could estimate $p(D|M_1)$ by drawing points from the prior on θ and calculating the mean likelihood.

Calculating $p(D|M)$

An analogy: the density of a particle represents the likelihood of a set of parameter values



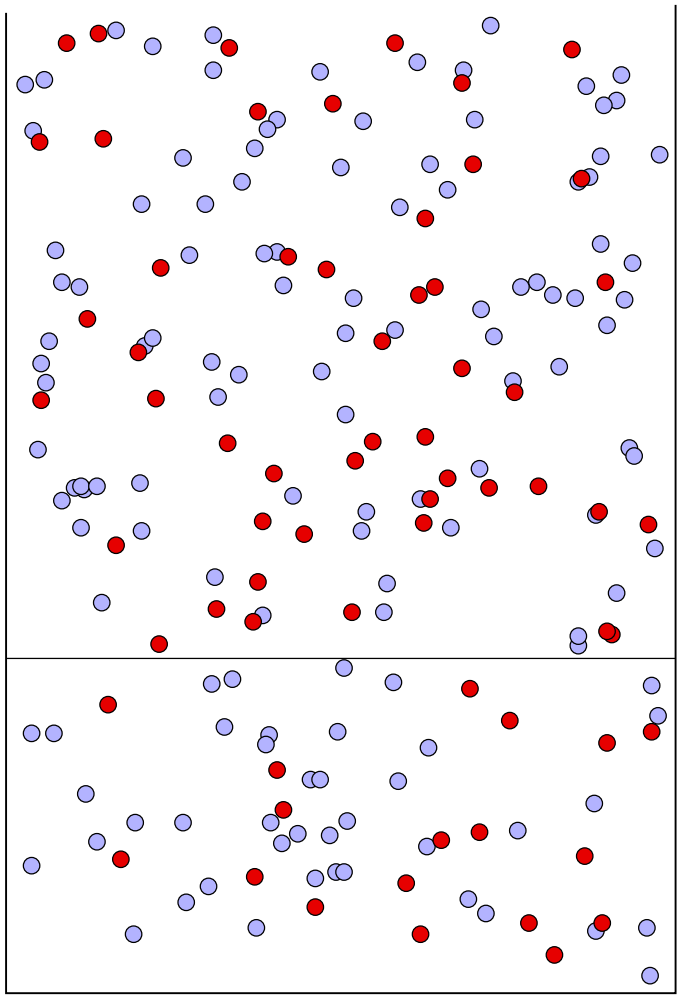
Consider particles suspended in syrup

$\frac{2}{3}$ of the particles are blue with mass $= 9$

$\frac{1}{3}$ of the particles are red with mass $= 18$

Mean mass in the population is:

$$\frac{2}{3} * 9 + \frac{1}{3} * 18 = 12$$

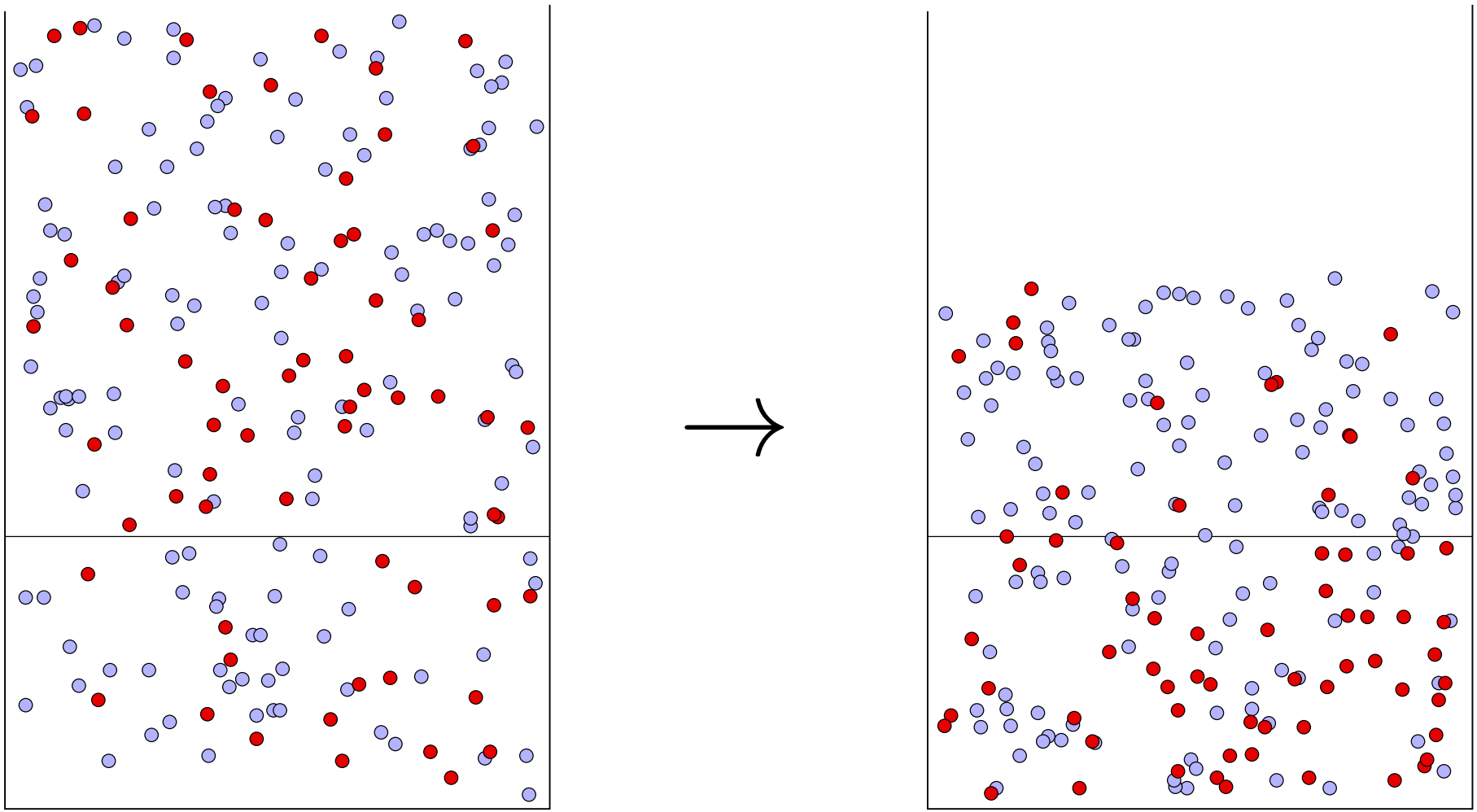


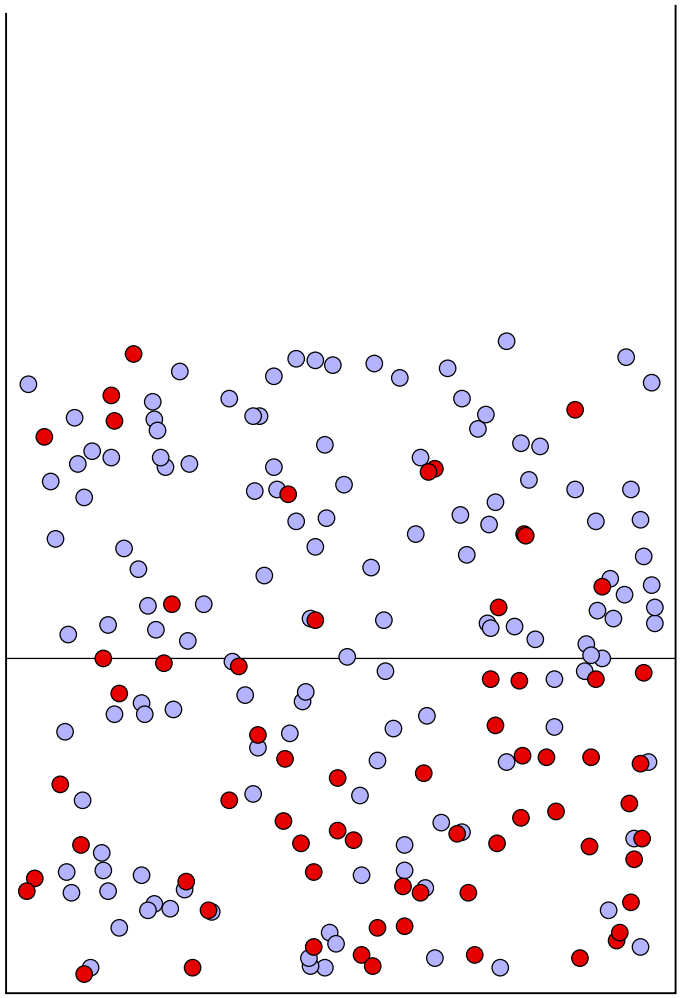
If everything is mixed well, then taking the mean mass from a sample gives a good estimate of the mean mass in the population

41 blue and 18 red in the bottom:

$$\frac{41*9+18*18}{59} \approx 11.75$$

Now, imagine a "Bayesian centrifuge" that enriches the bottom of the flask for heavy particles in exactly the same way that Bayesian MCMC prefers to sample points with high likelihood:





The arithmetic mean of our "centrifuged" sample is a horribly biased estimator of the mean mass:

52 blue and 52 red in the bottom

$$\frac{52*9+52*18}{104} = 13.5$$

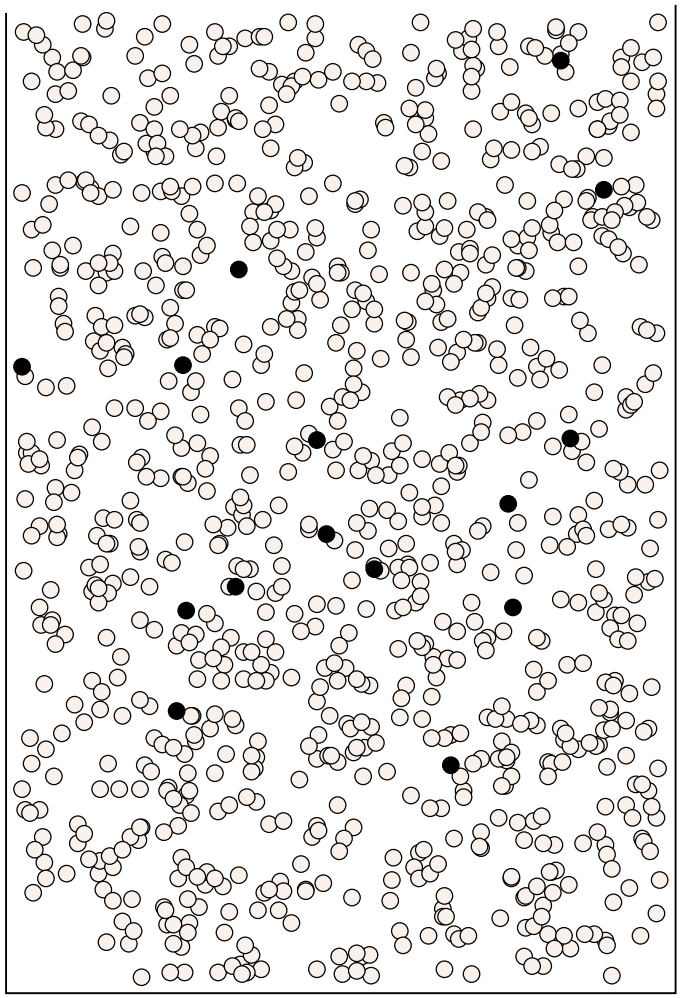
Interestingly, the harmonic mean of a centrifuged sample is good estimator:

$$\frac{104}{\frac{52}{9}+\frac{52}{18}} = 12$$

Harmonic mean estimator of the marginal likelihood

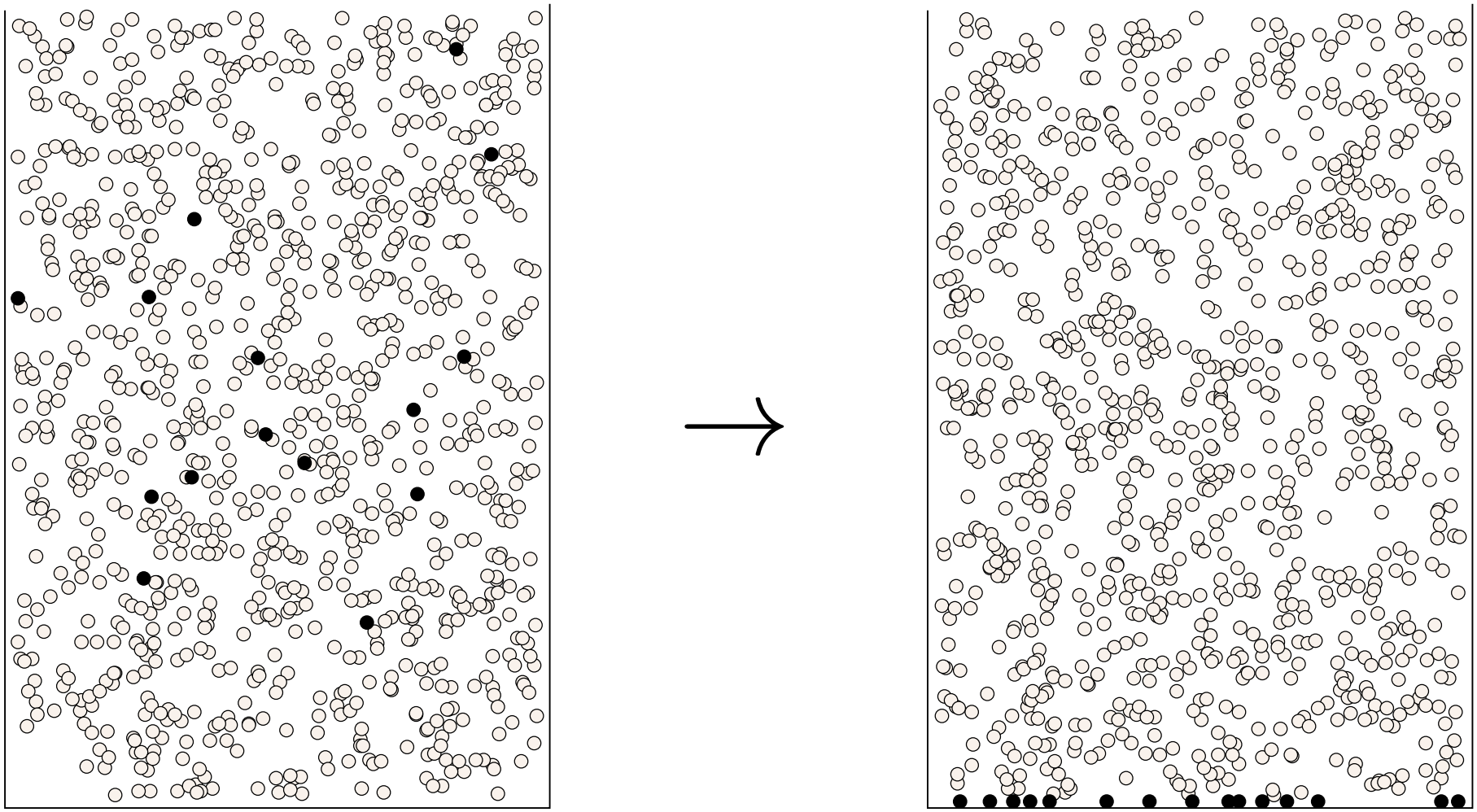
- appealing because it comes "for free" after we have sampled the posterior using MCMC,

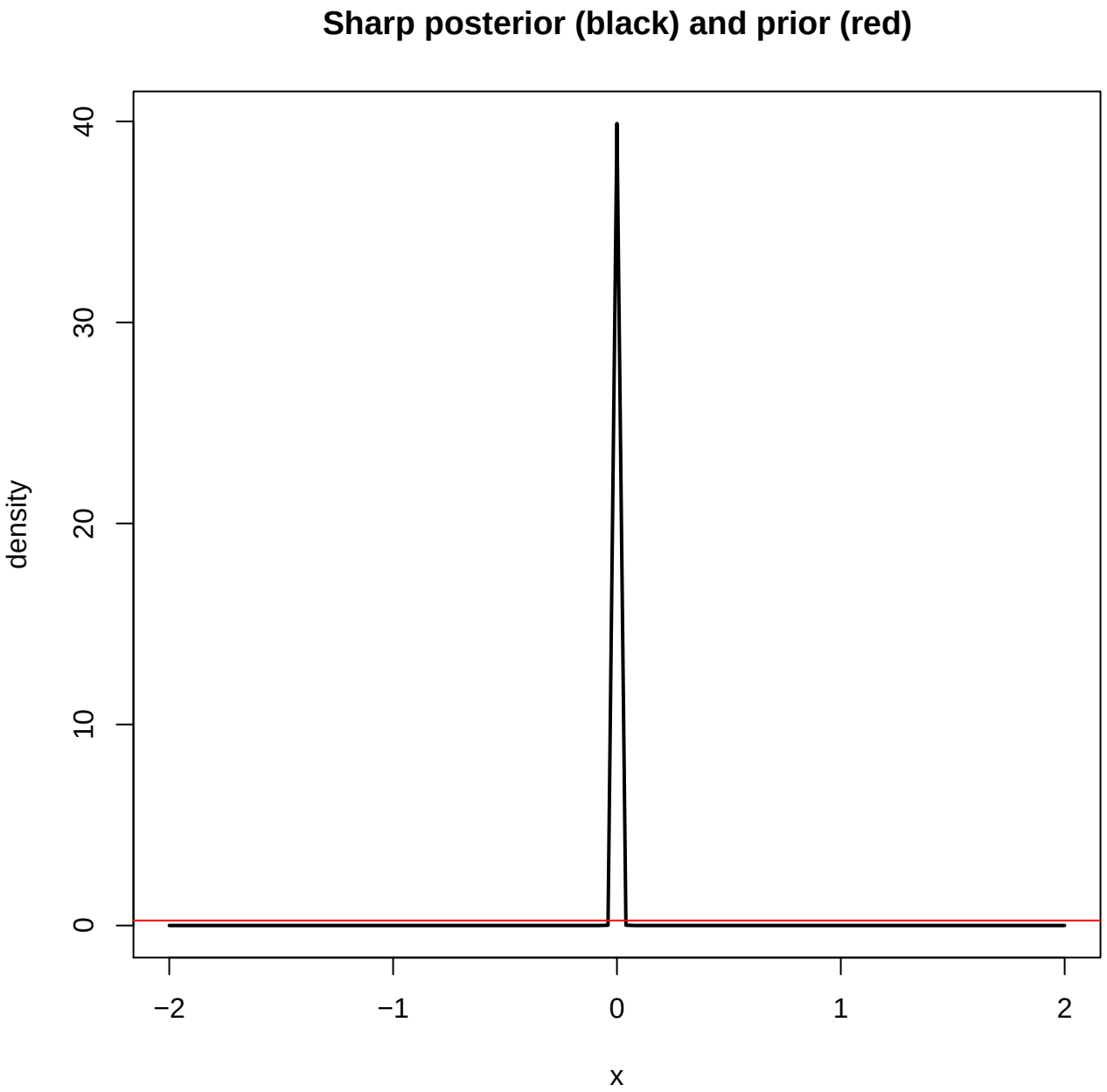
- unfortunately the estimator can have a huge variance associated with it in some (very common) cases. For example if:
  - the vast majority of parameter space has very low likelihood, and
  - a very small region has high likelihoods.



Think of a mixture in which most points (show in white) are only slightly denser than the syrup, and a few particles are made of lead.

Now, a moderate sample taken from bottom after our "Bayesian centrifuge" will probably result in a collection composed entirely of the heavy particles:



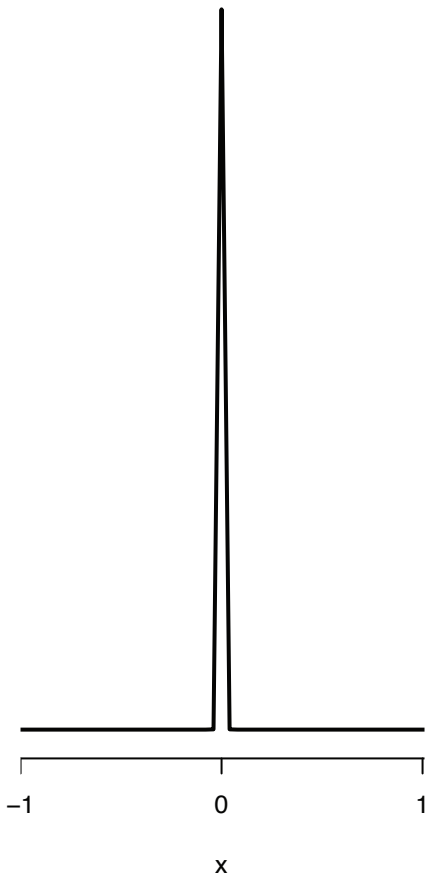

Sharp posterior (black) and prior (red)
density
0
10
20
30
40
−2
−1
0
1
2
x

From 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."

Two models with the same likelihood function, but different priors on the parameter (x). The result is marginal likelihoods that differ (by almost a factor of 2):



Steppingstone sampling

Xie, Lewis, Fan, Kuo, and Chen (Xie et al., 2011, accepted) and Fan, Wu, Chen, Kuo, and Lewis(Fan et al., 2010, in review) introduced a new method for estimating the marginal likelihood from several MCMC runs conducted with differing power posteriors densities:

$$p_\beta(\theta|D,M) = \frac{[p(D|\theta,M)p(\theta,M)]^\beta \, \pi(\theta)^{(1-\beta)}}{c_\beta}$$

where $0 \leq \beta \leq 1$ and c_β is a normalizing constant.

Intermediate values of β blend the posterior and the reference distribution. This is similar to the "heating" in MCMCMC.

Steppingstone sampling

When $\beta = 1$:

$$p_\beta(\theta|D, M) = \frac{[p(D|\theta, M)p(\theta, M)]^\beta \, \pi(\theta)^{(1-\beta)}}{c_\beta}$$

$$p_1(\theta|D, M) = \frac{[p(D|\theta, M)p(\theta, M)]}{c_1}$$

$p_1(\theta|D)$ is the posterior.

$c_1 = p(D|M)$, in other words, c_1 is the marginal likelihood.

Steppingstone sampling

When $\beta = 0$:

$$p_\beta(\theta|D, M) = \frac{[p(D|\theta, M)p(\theta, M)]^\beta \, \pi(\theta)^{(1-\beta)}}{c_\beta}$$

$$p_0(\theta|D, M) = \frac{\pi(\theta)}{c_0}$$

$\pi(\theta)$ is the reference distribution.

If we choose some analytically tractable reference distribution, $\pi(\theta)$, then we can calculate the density exactly and $c_0 = 1$.

Steppingstone sampling

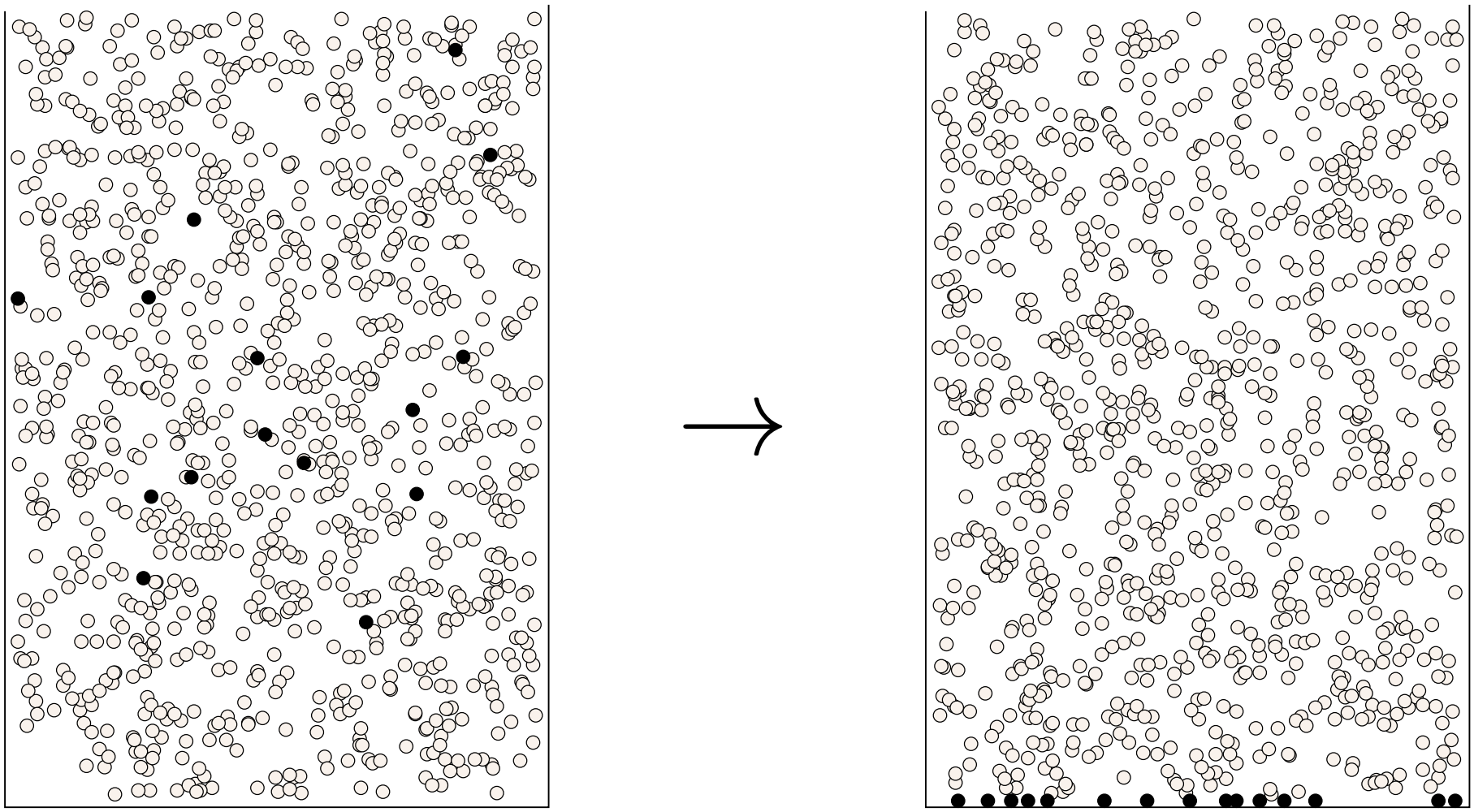
$$
\begin{aligned}
p(D|M) &= \frac{p(D|M)}{1.0} \\
&= \frac{c_1}{c_0} \\
&= \left(\frac{c_1}{c_{0.38}}\right)\left(\frac{c_{0.38}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.01}}\right)\left(\frac{c_{0.01}}{c_0}\right) \\
&= \left(\frac{c_1}{\cancel{c_{0.38}}}\right)\left(\frac{\cancel{c_{0.38}}}{\cancel{c_{0.1}}}\right)\left(\frac{\cancel{c_{0.1}}}{\cancel{c_{0.01}}}\right)\left(\frac{\cancel{c_{0.01}}}{c_0}\right)
\end{aligned}
$$

This corresponds to using β of $\{0, 0.01, 0.1, 0.38, 1\}$ as stepping stones between the reference distribution and the posterior.

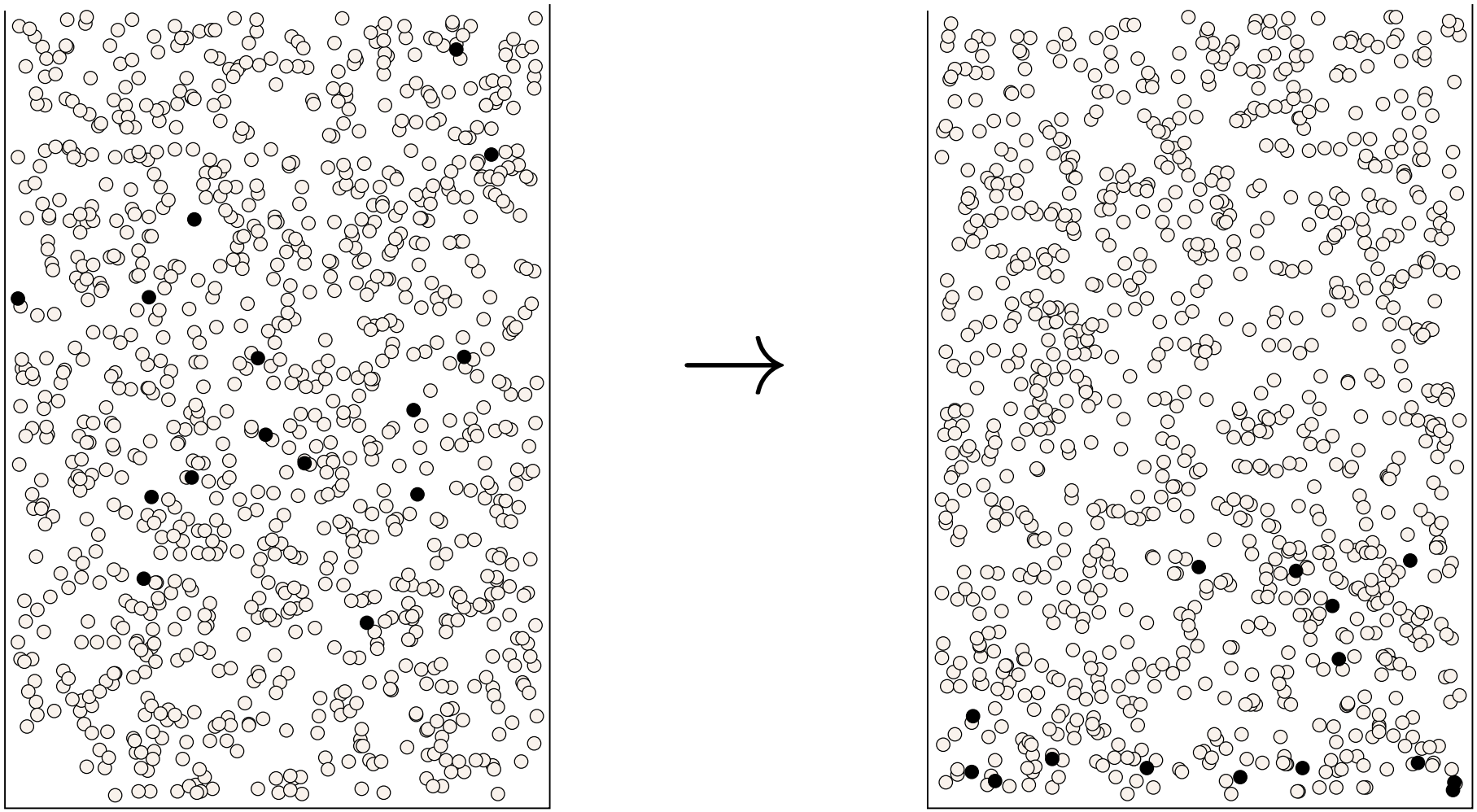
$$p(D|M) = \frac{c_1}{c_0} = \left(\frac{c_1}{c_{0.38}}\right)\left(\frac{c_{0.38}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.01}}\right)\left(\frac{c_{0.01}}{c_0}\right)$$

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

Back to the centrifuge example: The "steppingstones" correspond to different speeds of our centrifuge and repeat the experiment. If we only change the speed a bit we'll get a similar result:



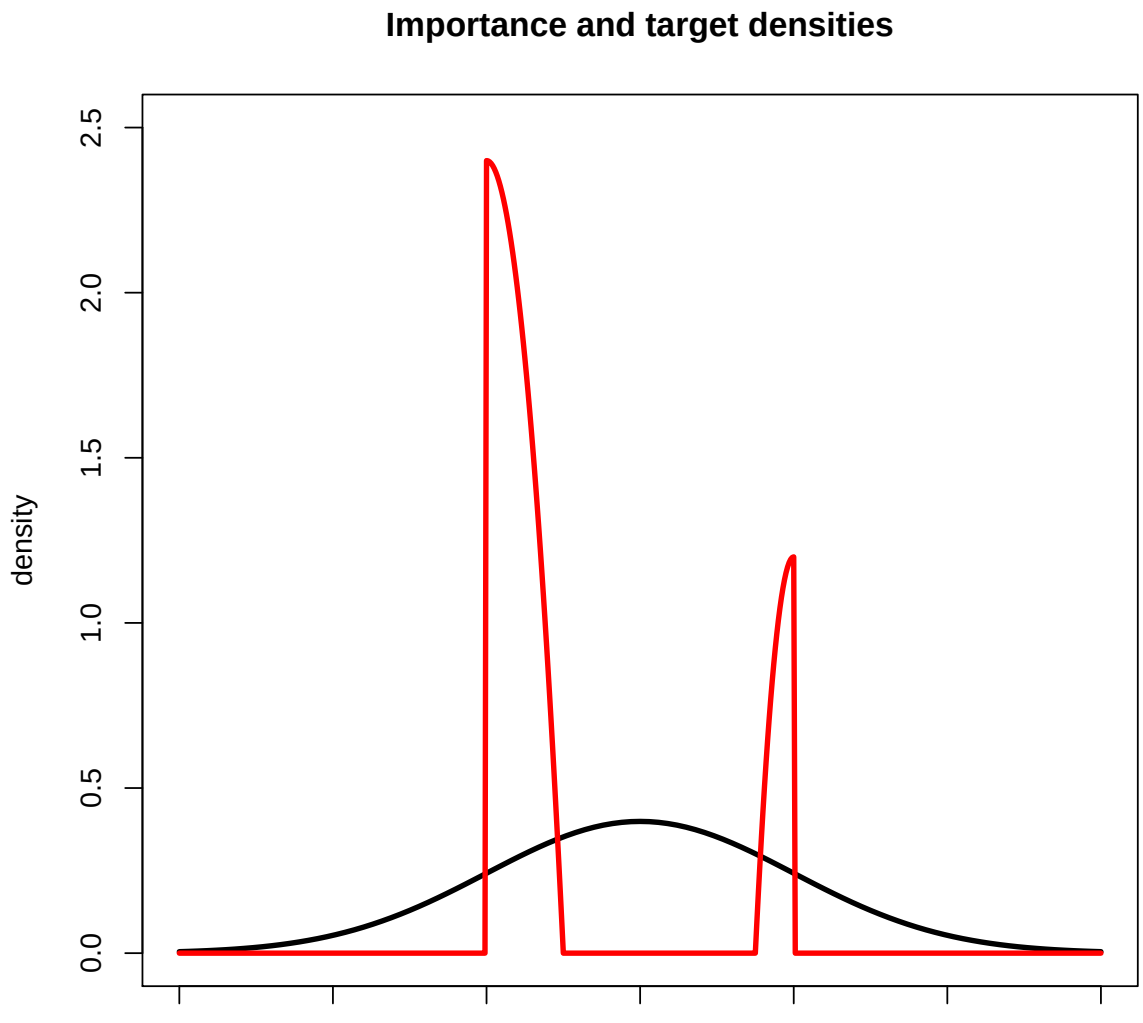
By repeating the experiment with lower and lower speeds of the centrifuge we can eventually find a sample that gives us a good mixture of heavy and light particles:

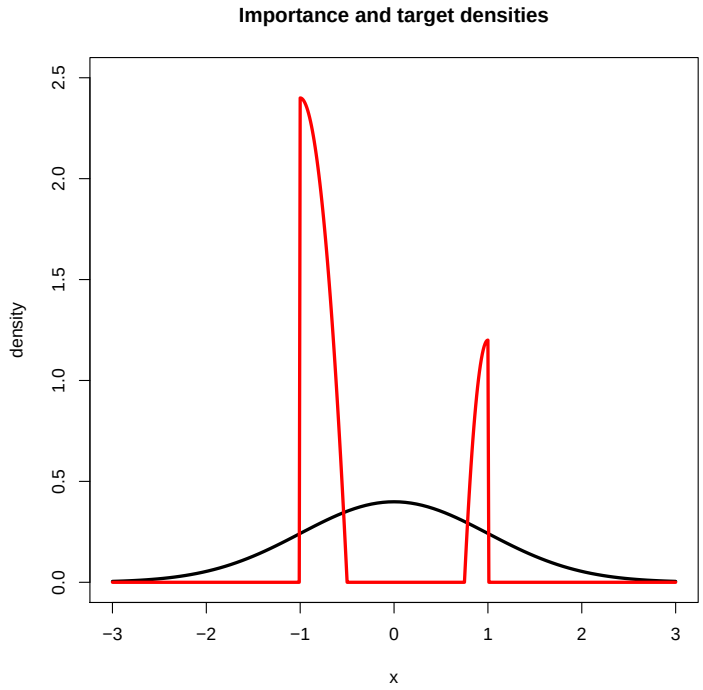


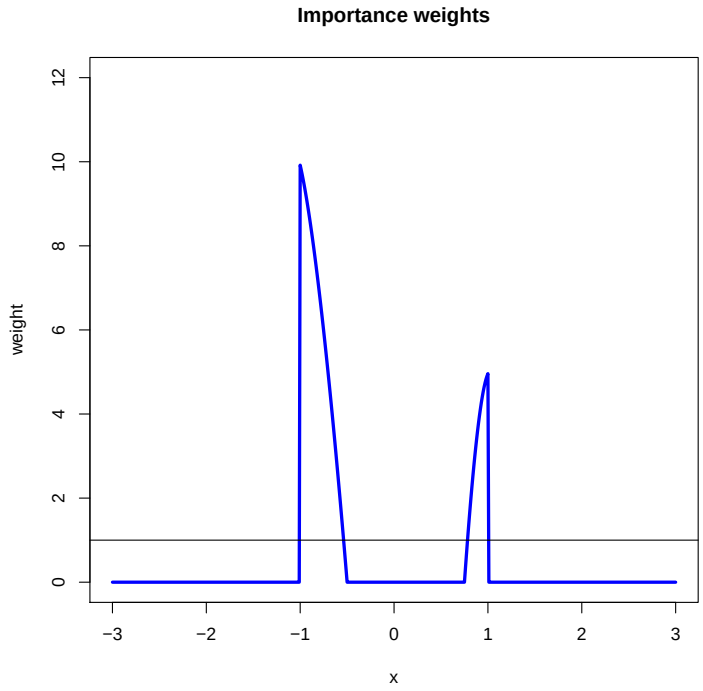
Using information from all of our experiments we can estimate the mean mass accurately.

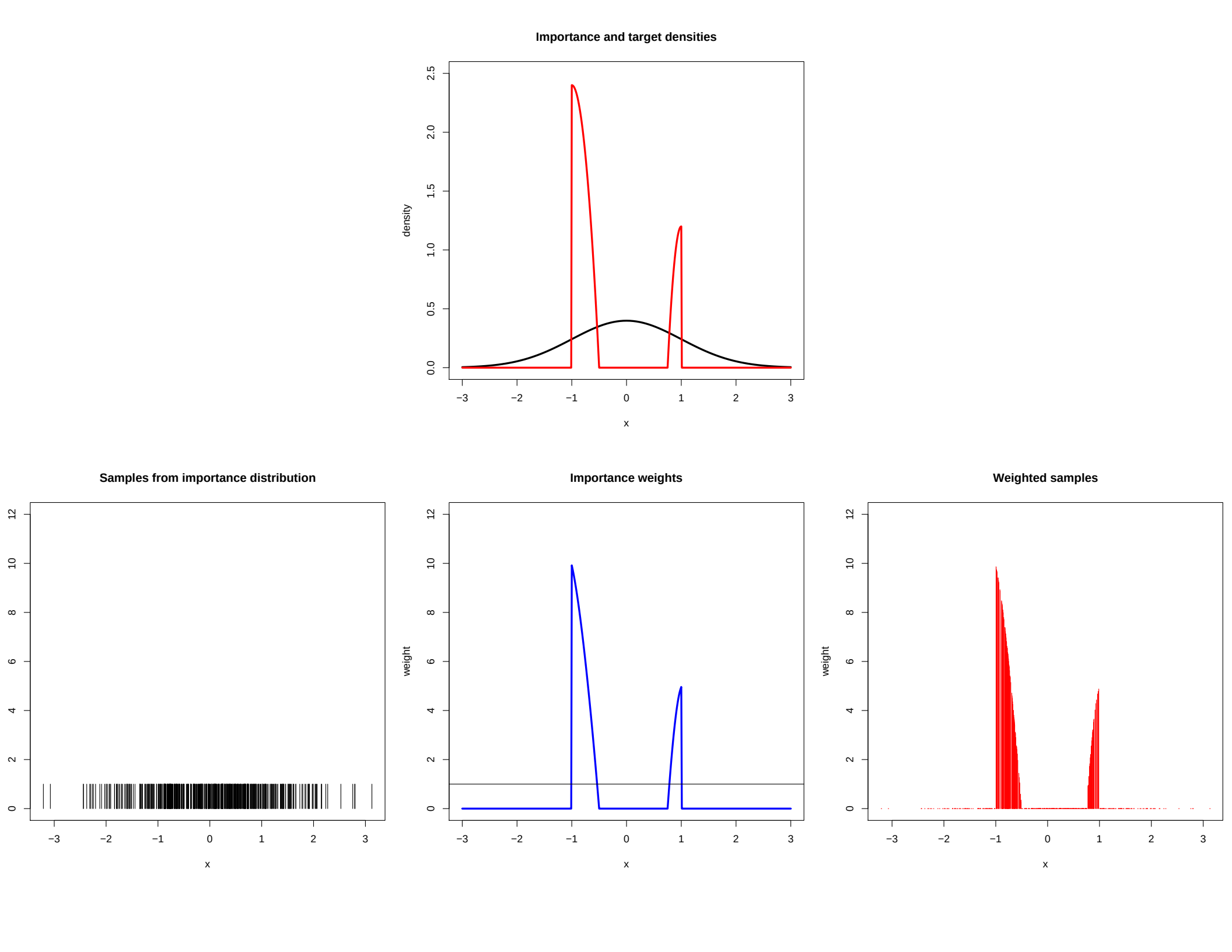
We can use a technique called importance sampling to estimate the ratio of the normalizing constants between adjacent steppingstones.

Importance sampling: we simulate points from one distribution, and then reweight the points to transform them into samples from a target distribution that we are interested in:




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
x

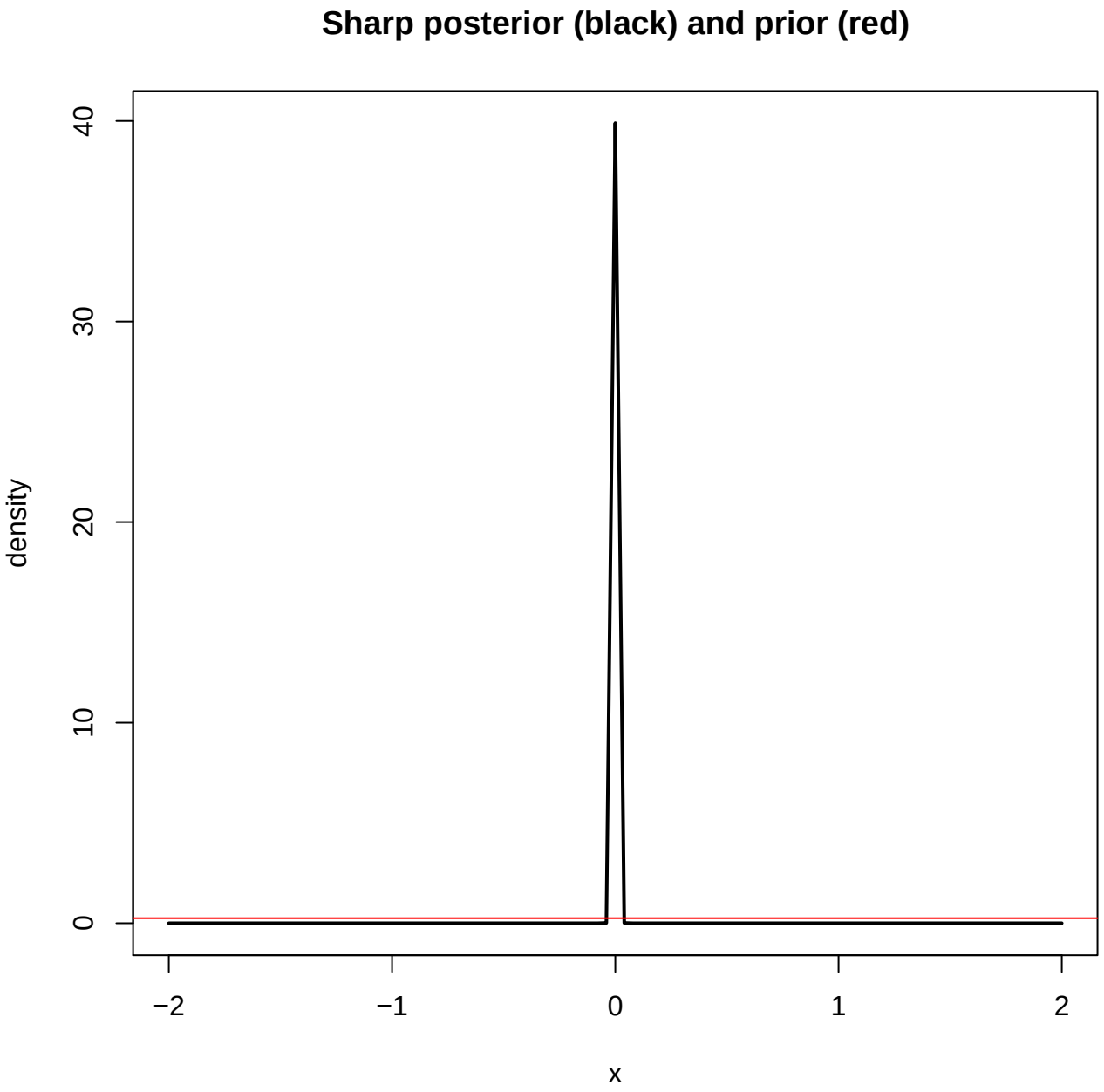

Importance and target densities
density
x
Samples from importance distribution
Importance weights
weight
Weighted samples

Importance sampling

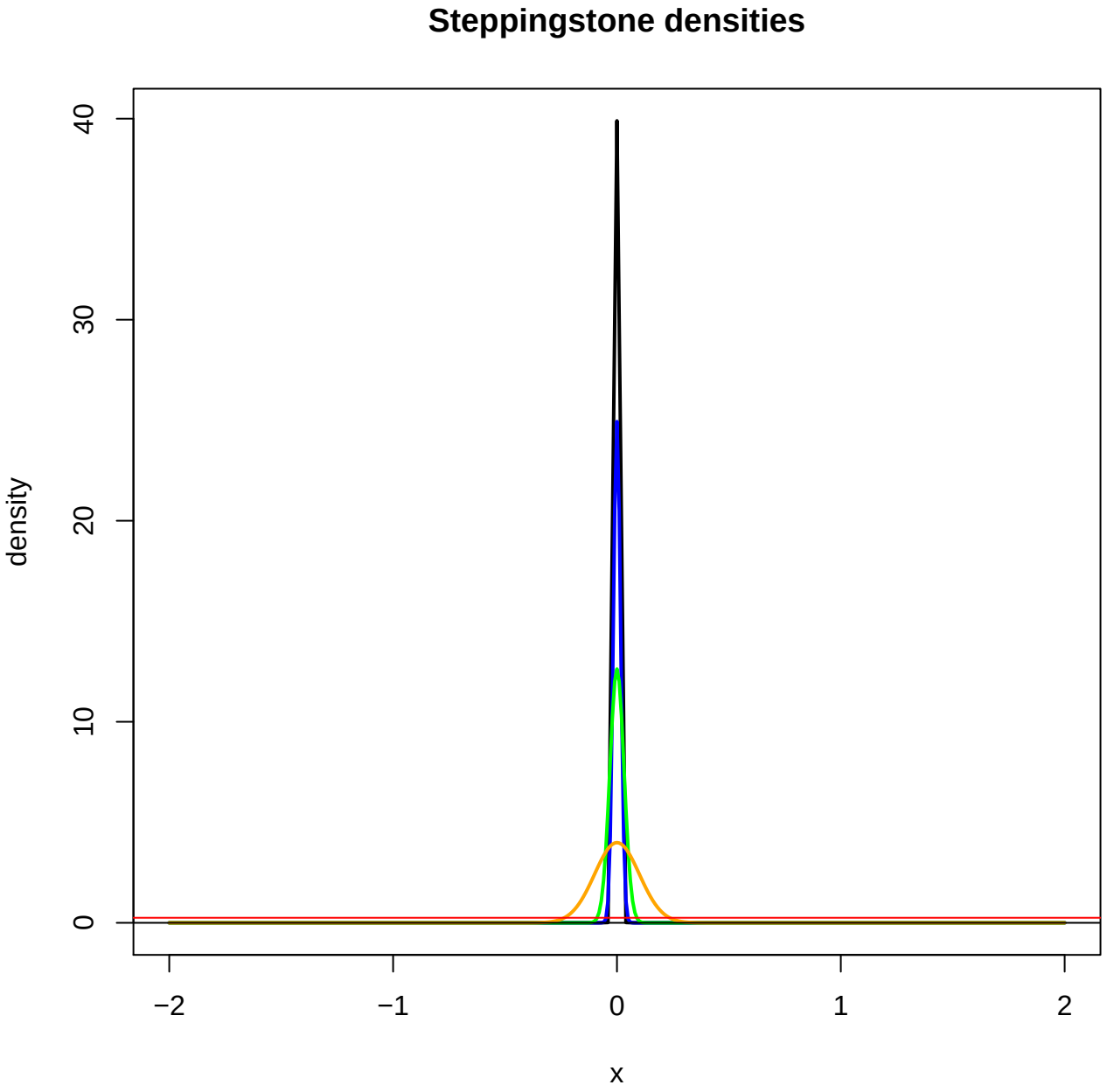
The method works well if the importance distribution is:

- 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:



The key to steppingstone sampling is to use a slightly vaguer distributions as importance distributions in a series of steps:



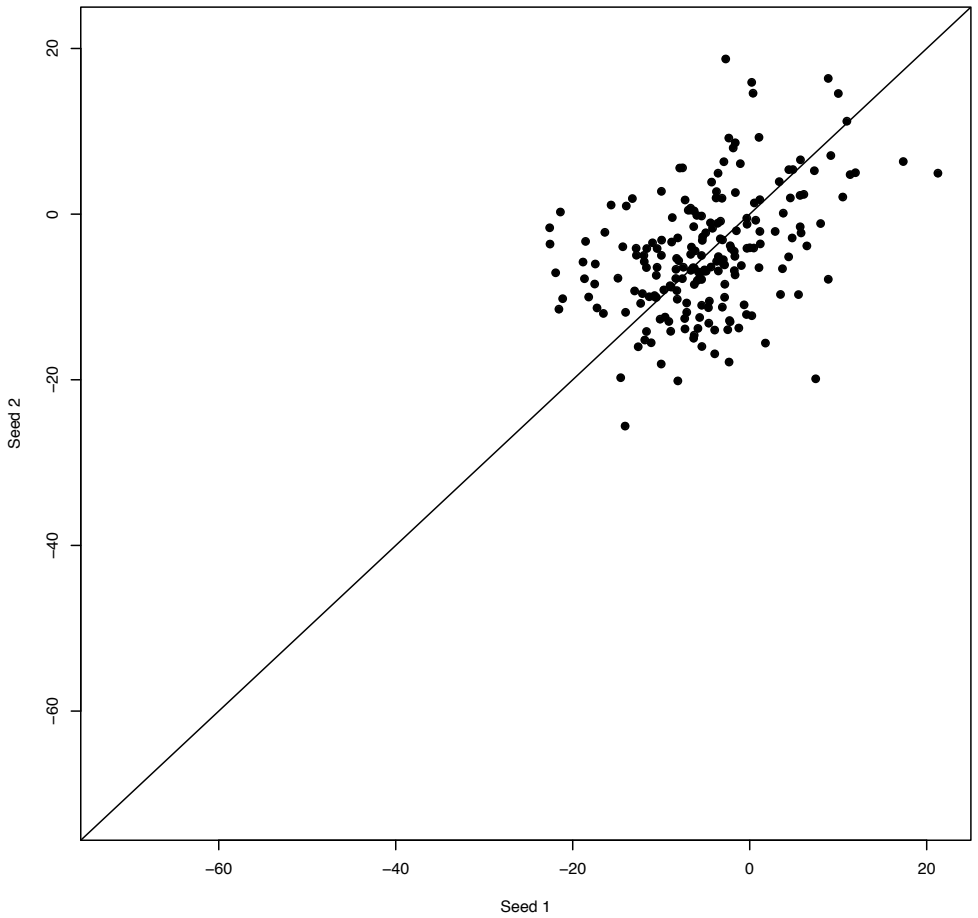
Path sampling

Lartillot and Philippe (2006) introduced two forms of path sampling (which they called "thermodynamic integration"), one of which is similar to steppingstone sampling in that it morphs the prior into the posterior through power posterior distributions.

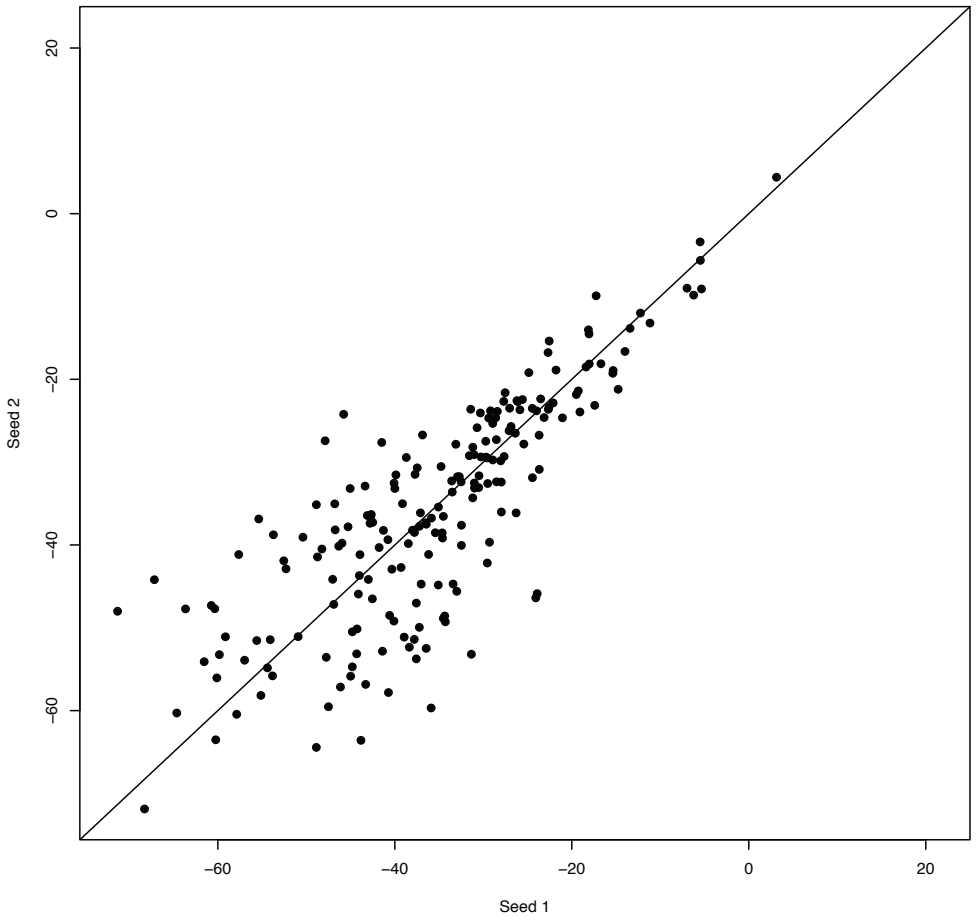
Phycas also implements a discrete step version of Lartillot and Philippe's "annealing-melting integration."

The Bayes Factor for a complex model compared to a simple (true) model. (estimated twice to assess repeatability)

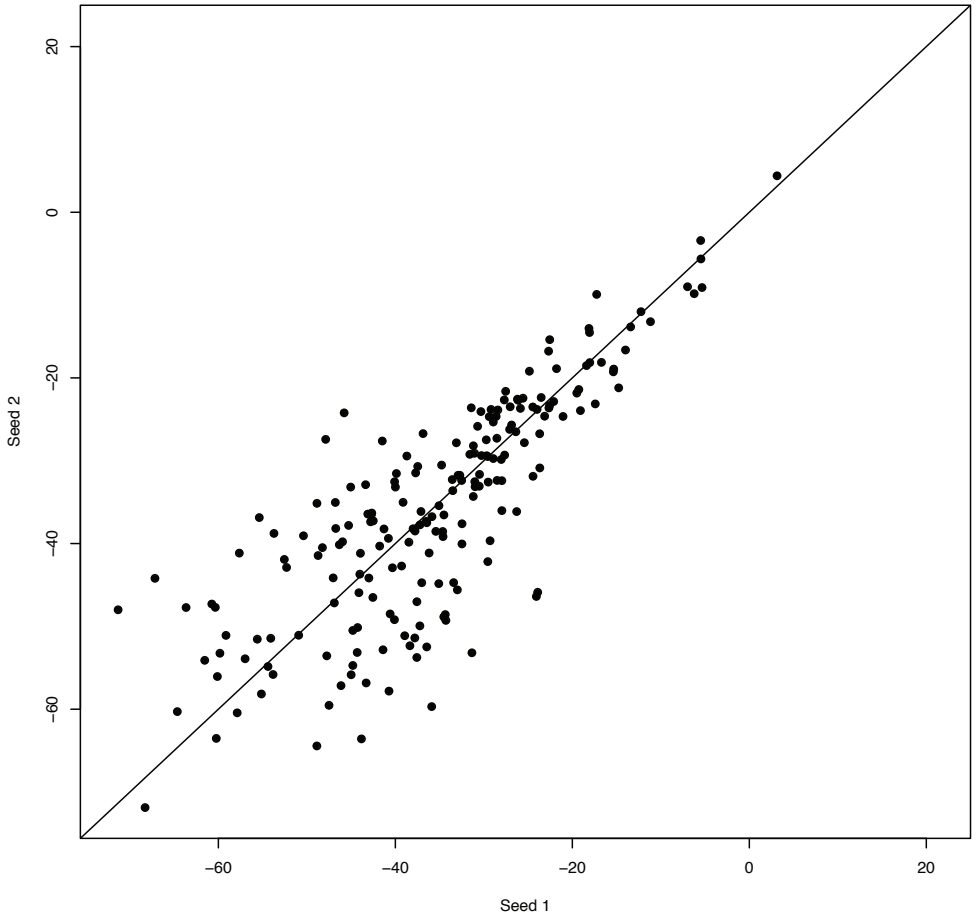
Harmonic mean



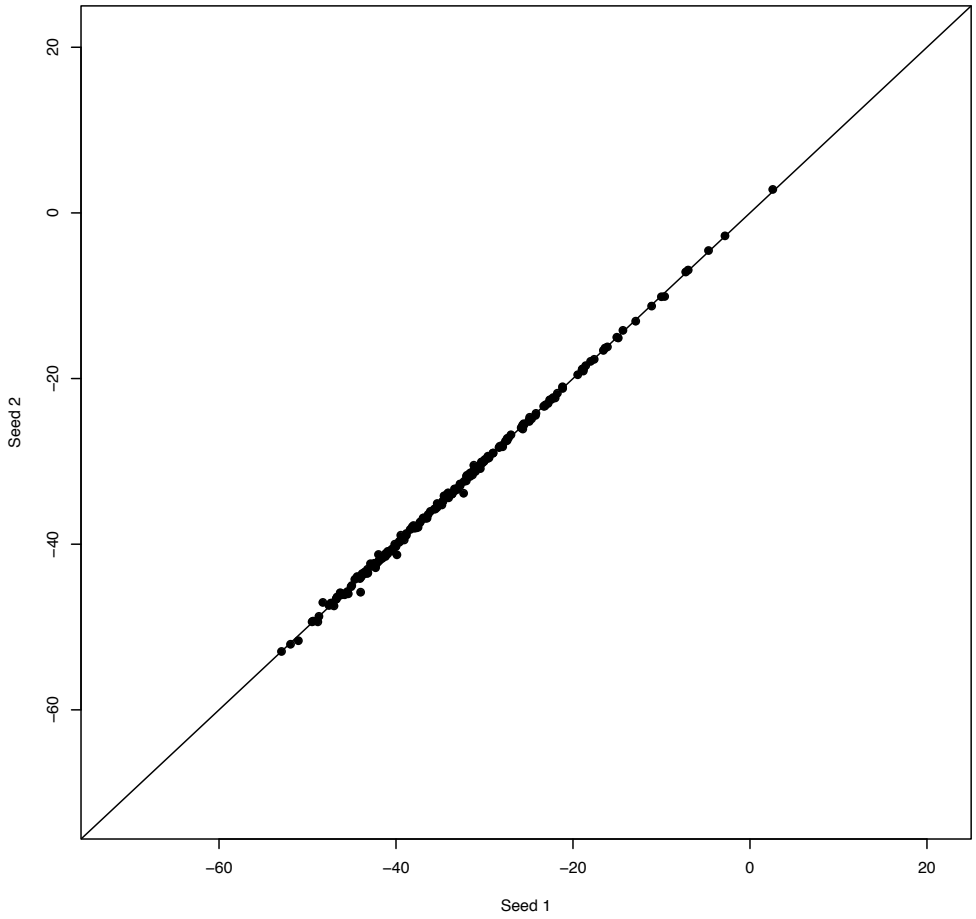
Original Steppingstone (Xie et al., 2011): (blends posterior and prior)



Original Steppingstone (Xie et al., 2011):
(blends posterior and prior)



New Steppingstone (Fan et al., 2010):
(blends posterior and a simple version of the posterior fit using tractable distributions)



References

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2010). Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, 28(1):523–532.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160.