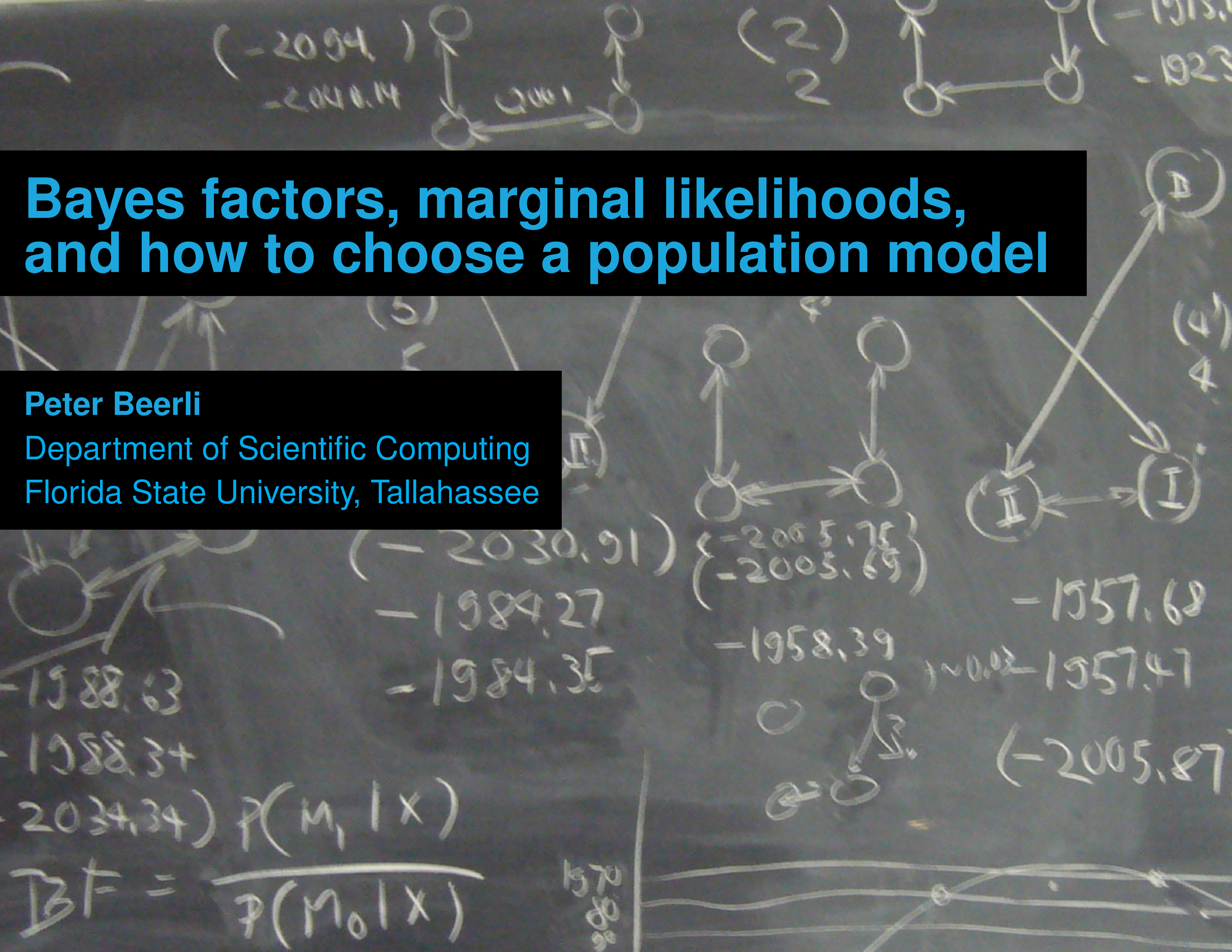


Bayes factors, marginal likelihoods, and how to choose a population model

Peter Beerli
Department of Scientific Computing
Florida State University, Tallahassee

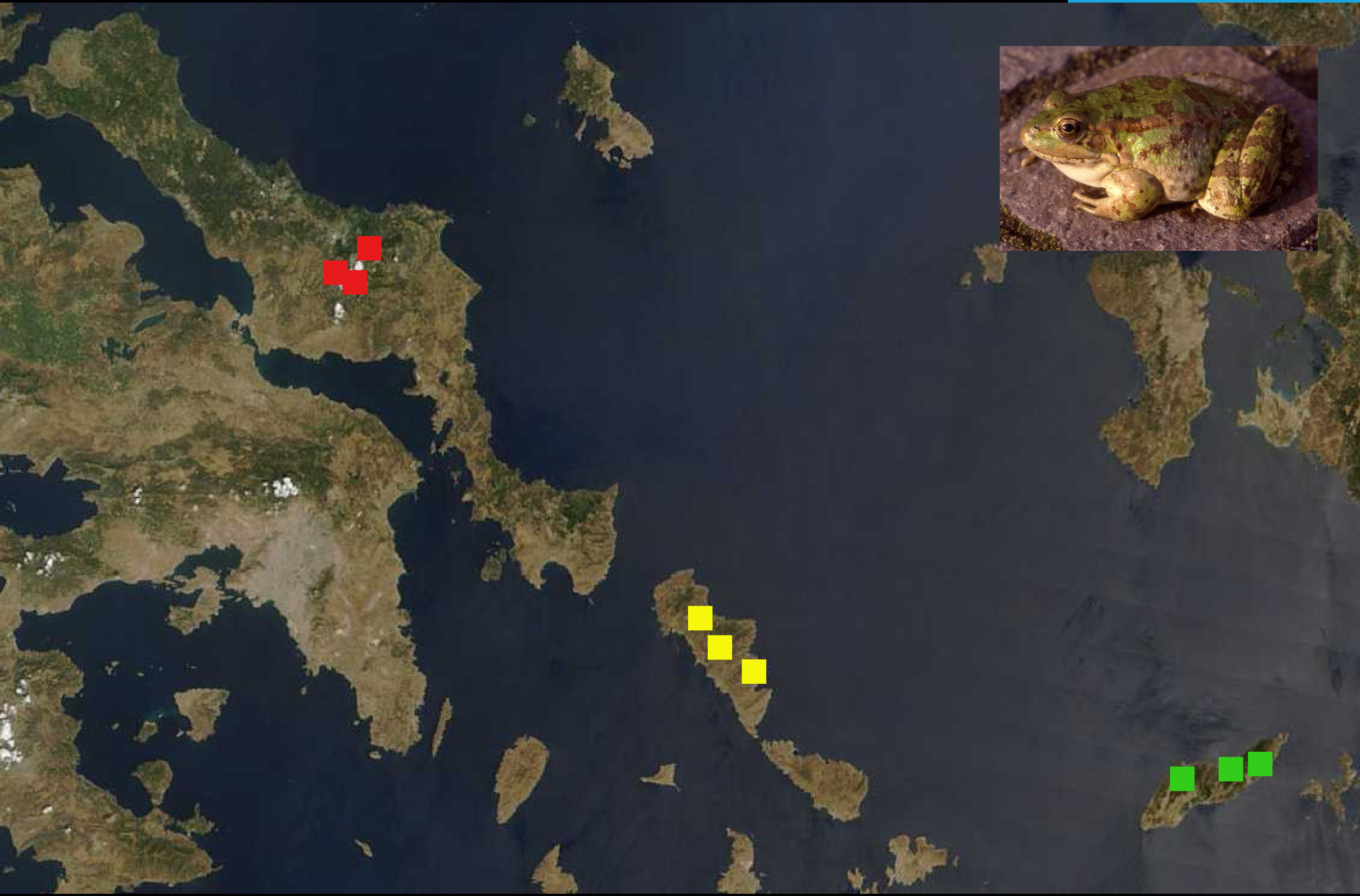
Overview

1. Location versus Population
2. Bayes factors, what are they and how to calculate them
3. Marginal likelihoods, what are they and how to calculate them
4. Examples: simulated and real data
5. Resources: replicated runs, cluster computing

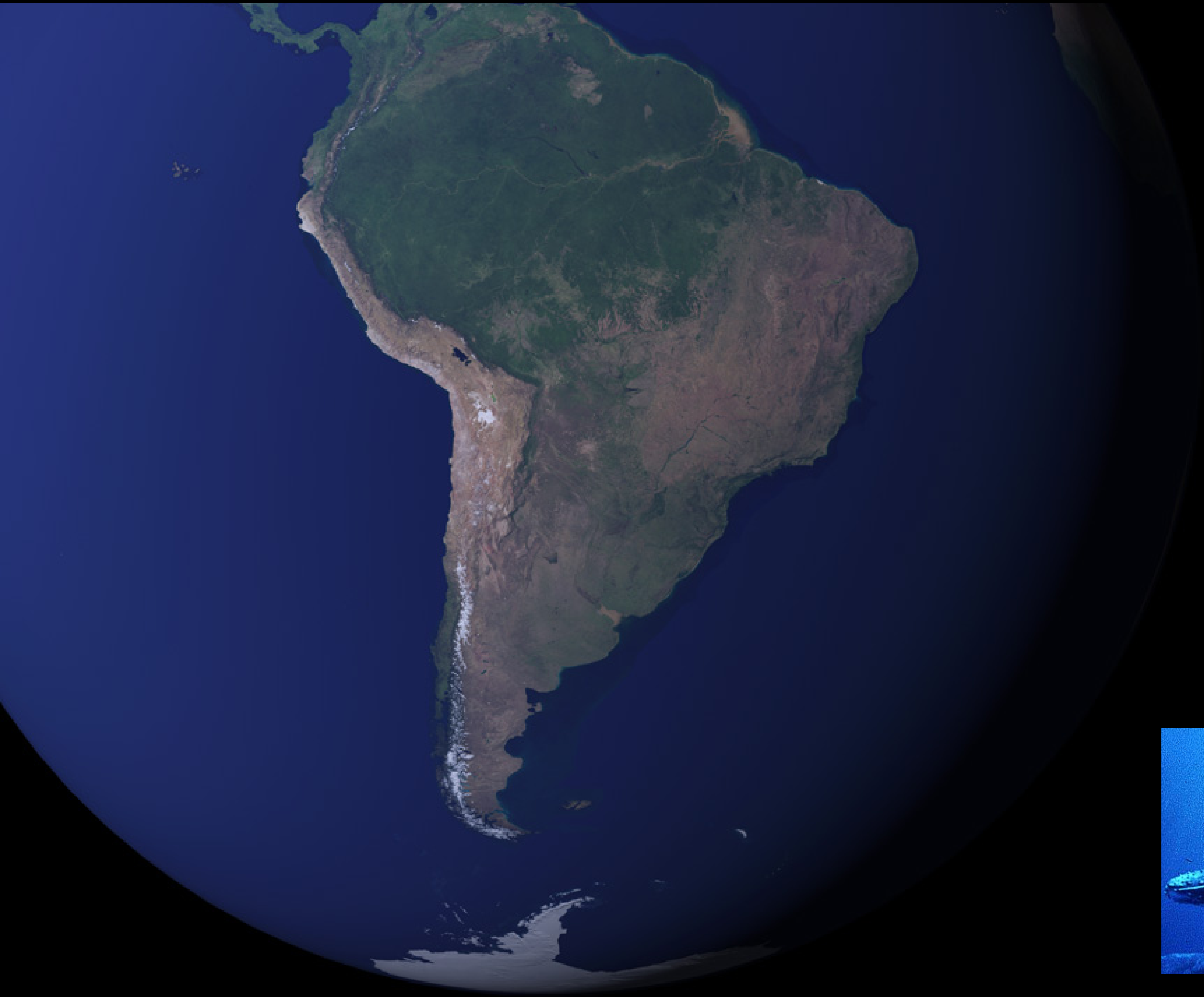
Location versus Population

Location versus Population

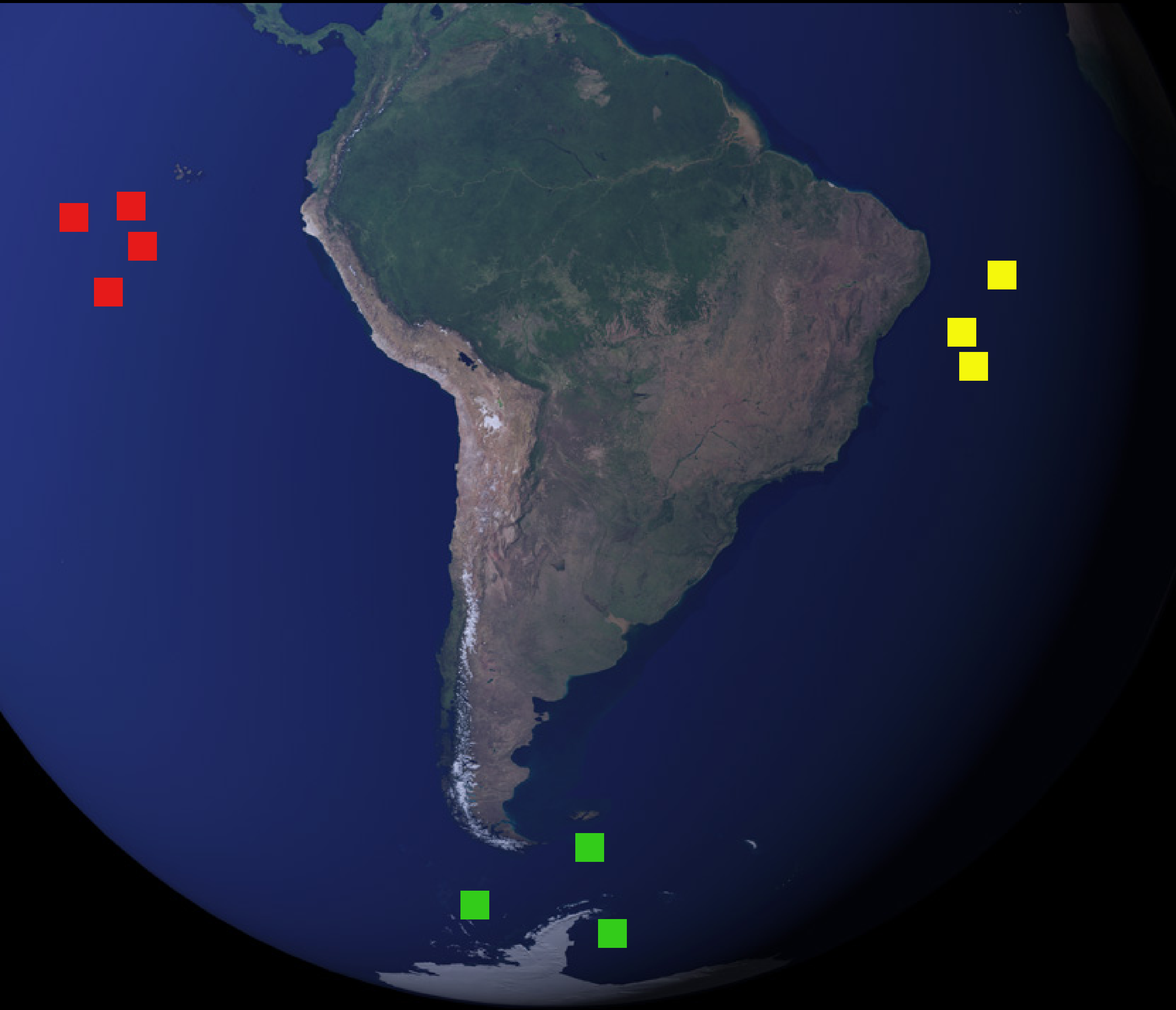
Location ≈ Population



Location versus Population



Location $\overset{?}{=}$ Population



Model comparison

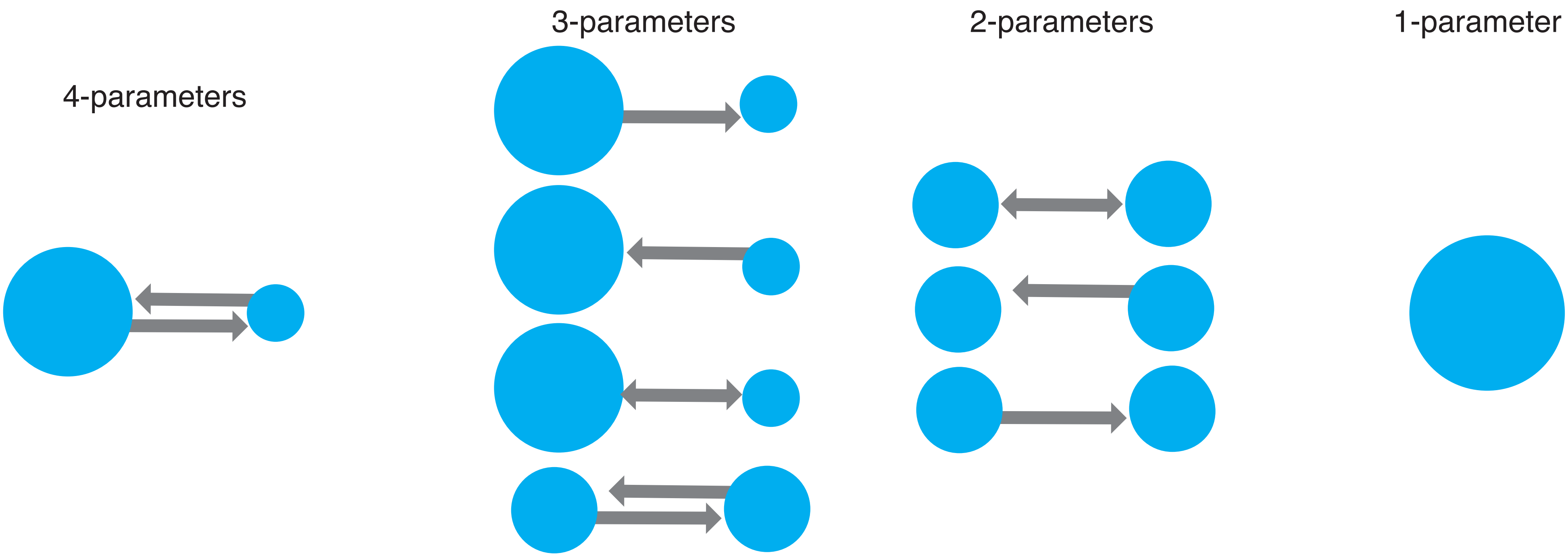
- Several tests that establish whether two locations belong to the same population exist. The test by Hudson and Kaplan (1995) seemed particularly powerful even with a single locus.

- These days researchers mostly use the program STRUCTURE to establish the number of populations.

- A procedure that not only can handle panmixia versus all other gene flow models would help.

Model comparison

For example we want to compare some of these models



4-parameters

3-parameters

2-parameters

1-parameter

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.

Kass and Raftery (1995) popularized the Bayes Factor as a Bayesian alternative to the LRT.

Bayes factor

In a Bayesian context we could look at the posterior odds ratio or equivalently the Bayes factors.

$$\mathrm{p}(\mathrm{M}_1|\mathrm{X}) = \frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}$$

$$\frac{\mathrm{p}(\mathrm{M}_1|\mathrm{X})}{\mathrm{p}(\mathrm{M}_2|\mathrm{X})} = \frac{\mathrm{p}(\mathrm{M}_1)}{\mathrm{p}(\mathrm{M}_2)} \times \frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)}$$

$$\mathrm{BF} = \frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} \qquad \mathrm{LBF} = 2\ln\mathrm{BF} = 2\ln\left(\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)}\right)$$

The magnitude of BF gives us evidence against hypothesis M_2

$$\mathrm{LBF} = 2\ln\mathrm{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

Marginal likelihood

So why are we not all running BF analyses instead of the AIC, BIC, LRT?

Typically, it is rather difficult to calculate the marginal likelihoods with good accuracy, because most often we only approximate the posterior distribution using Markov chain Monte Carlo (MCMC).
In MCMC we need to know only differences and therefore we typically do not need to calculate the denominator to calculate the Posterior distribution $\mathrm{p}(\Theta|\mathrm{X})$:

$$\mathrm{p}(\Theta|\mathrm{X},\mathrm{M}) = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\mathrm{p}(\mathrm{X}|\mathrm{M})} = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\int_\Theta \mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)\mathrm{d}\Theta}$$

where $\mathrm{p}(\mathrm{X}|\mathrm{M})$ is the marginal likelihood.

Harmonic mean estimator

[Common approximation, used in programs such a MrBayes and Beast]
The harmonic mean estimator applied to our specific problem can be described using an importance sampling approach

$$\mathrm{p(X|M)} = \frac{\int_G \mathrm{p(X|G, M_i)p(G)dG}}{\int_G \mathrm{p(G)dG}}$$

which is approximated after some shuffling wth expectations by

$$\mathrm{p(X|M)} \simeq \frac{1}{\frac{1}{n}\sum_j^n \frac{1}{\mathrm{p(X|G,M)}}}, \quad G_j \sim \mathrm{p(G|X, M)}.$$

$$\ell_H = \ln \mathrm{p(X|M)}$$

Harmonic mean estimator

[Common approximation, used in programs such a MrBayes and Beast]
The harmonic mean estimator applied to our specific problem can be described using an importance sampling approach

$$\mathrm{p(X|M)} = \frac{\int_G \mathrm{p(X|G,M_i)p(G)dG}}{\int_G \mathrm{p(G)dG}}$$

which is approximated after some shuffling with expectations by

$$\mathrm{p(X|M)} \simeq \frac{1}{\frac{1}{n}\sum_j^n \frac{1}{\mathrm{p(X|G,M)}}}, \quad G_j \sim \mathrm{p(G|X,M)}.$$

$$\ell_H = \ln \mathrm{p(X|M)}$$

The Harmonic Mean of the Likelihood:
Worst Monte Carlo Method Ever
Radford Neal 2008

Thermodynamic integration

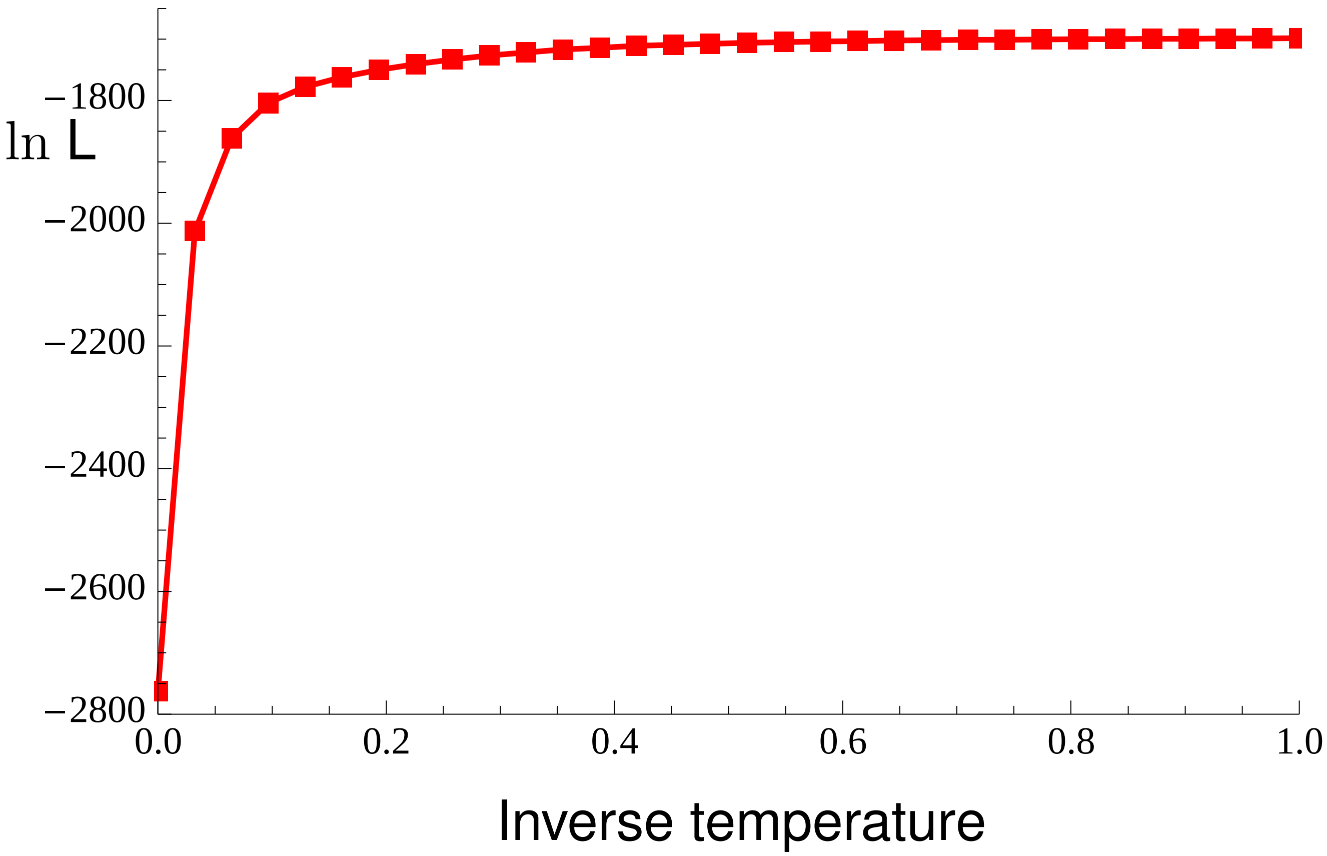
$$\ell_T = \ln \mathrm{p}(\mathrm{X}|\mathrm{M_i}) = \int_0^1 \mathbb{E}(\ln \mathrm{p_t}(\mathrm{X}|\mathrm{M_i}))\mathrm{dt}$$

which we approximate using the trapezoidal rule for $t_0 = 0 < t_1 < ... < t_n = 1$ using

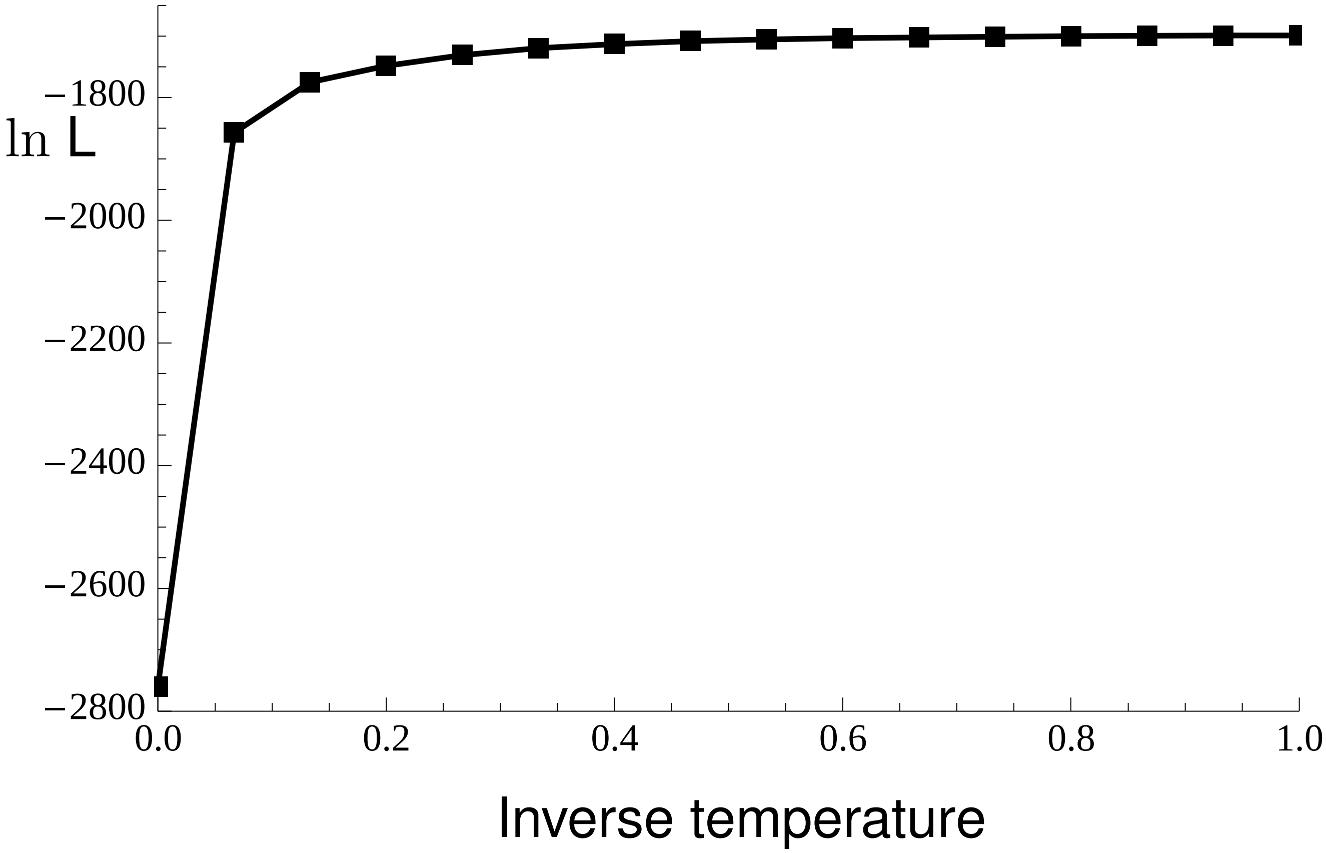
$$\mathbb{E}(\ln \mathrm{p_t}(\mathrm{X}|\mathrm{M_i})) \approx \frac{1}{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{m}} \ln \mathrm{p_{t_z}}(\mathrm{X}|\mathrm{G_j}, \mathrm{M_i})$$

Path sampling: Gelman and Meng (1998), Friel and Pettitt (2007,2009)
Phylogeny: Lartillot and Phillipe (2006),
Wu et al (2011), Xie et al (2011) [Paul Lewis]
Population genetics: Beerli and Palczewski 2010

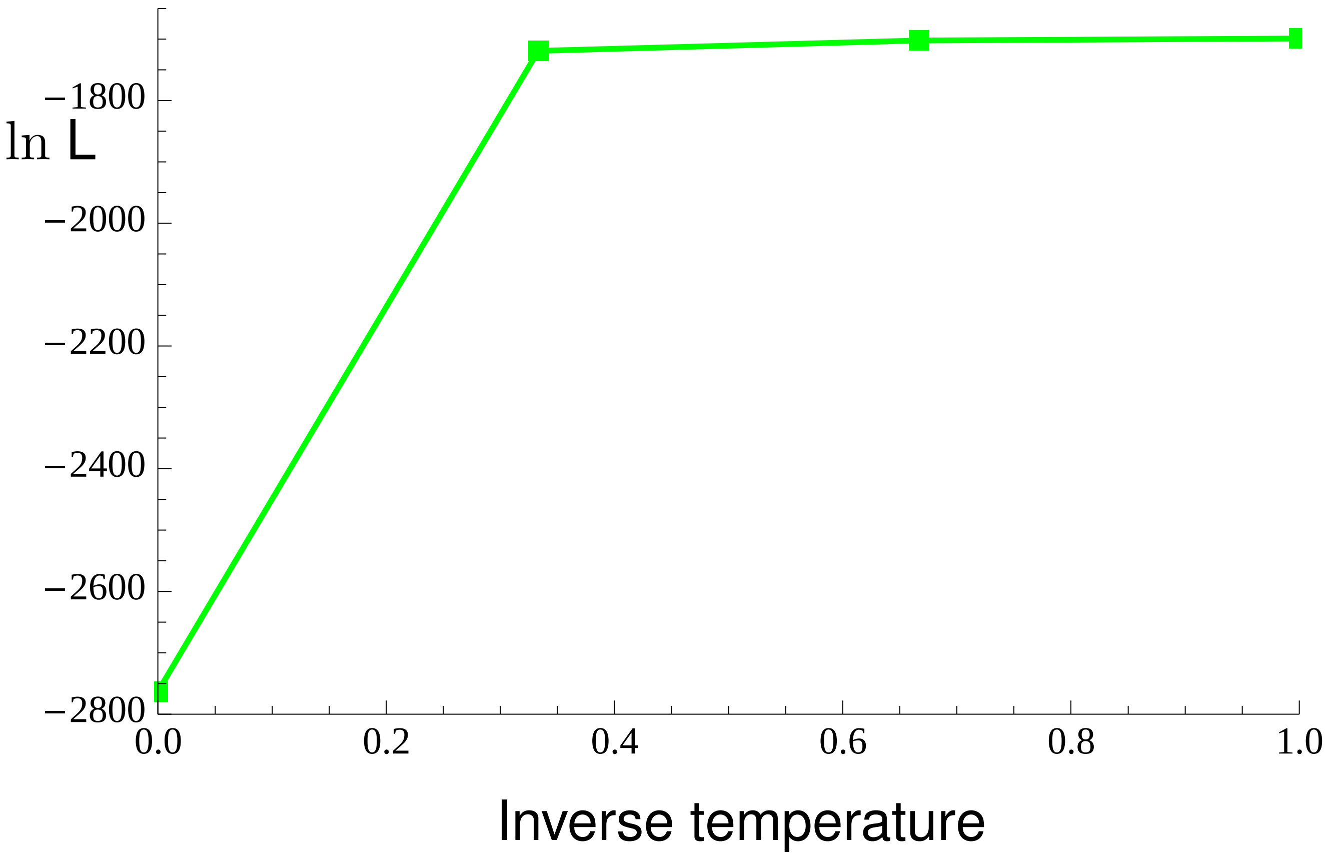
Thermodynamic integration



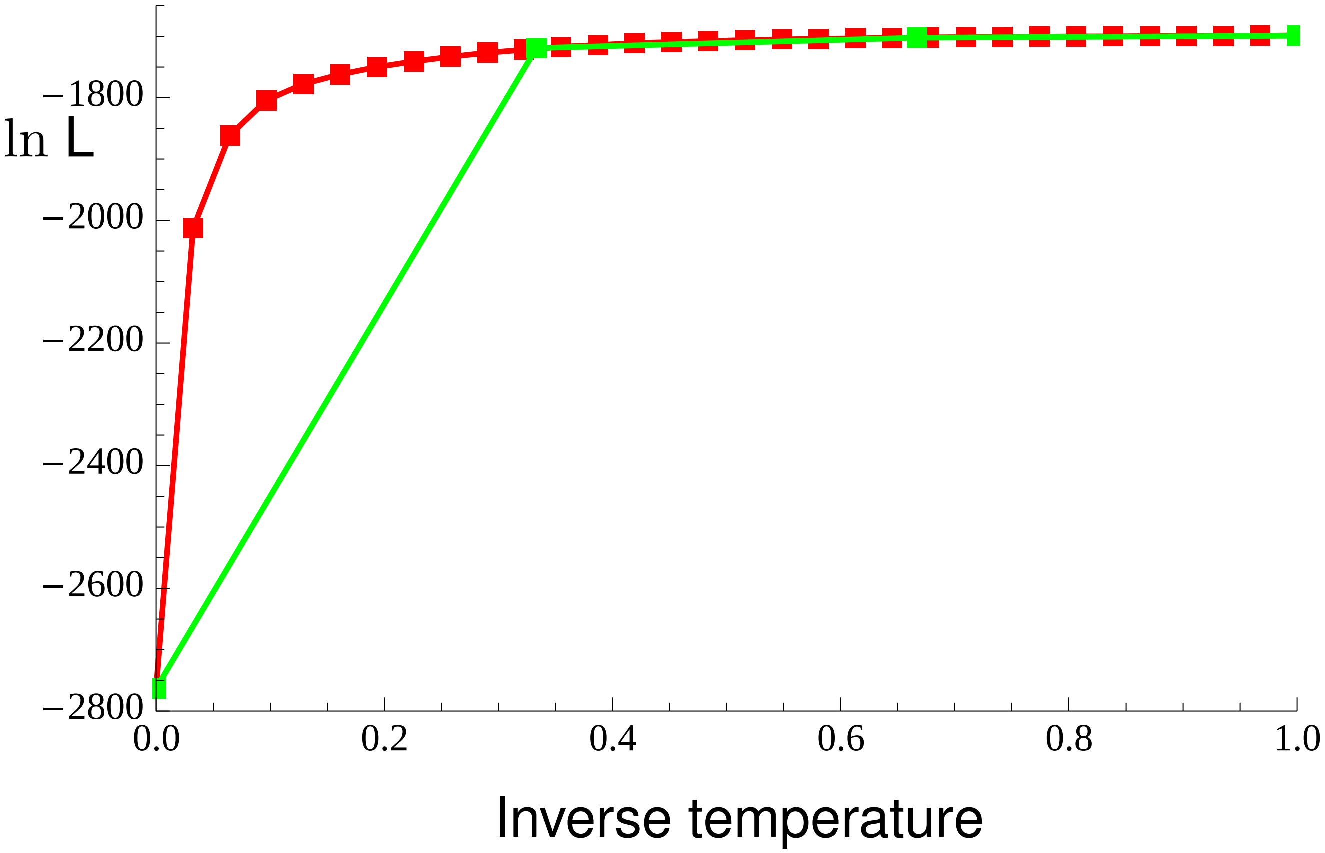
Thermodynamic integration



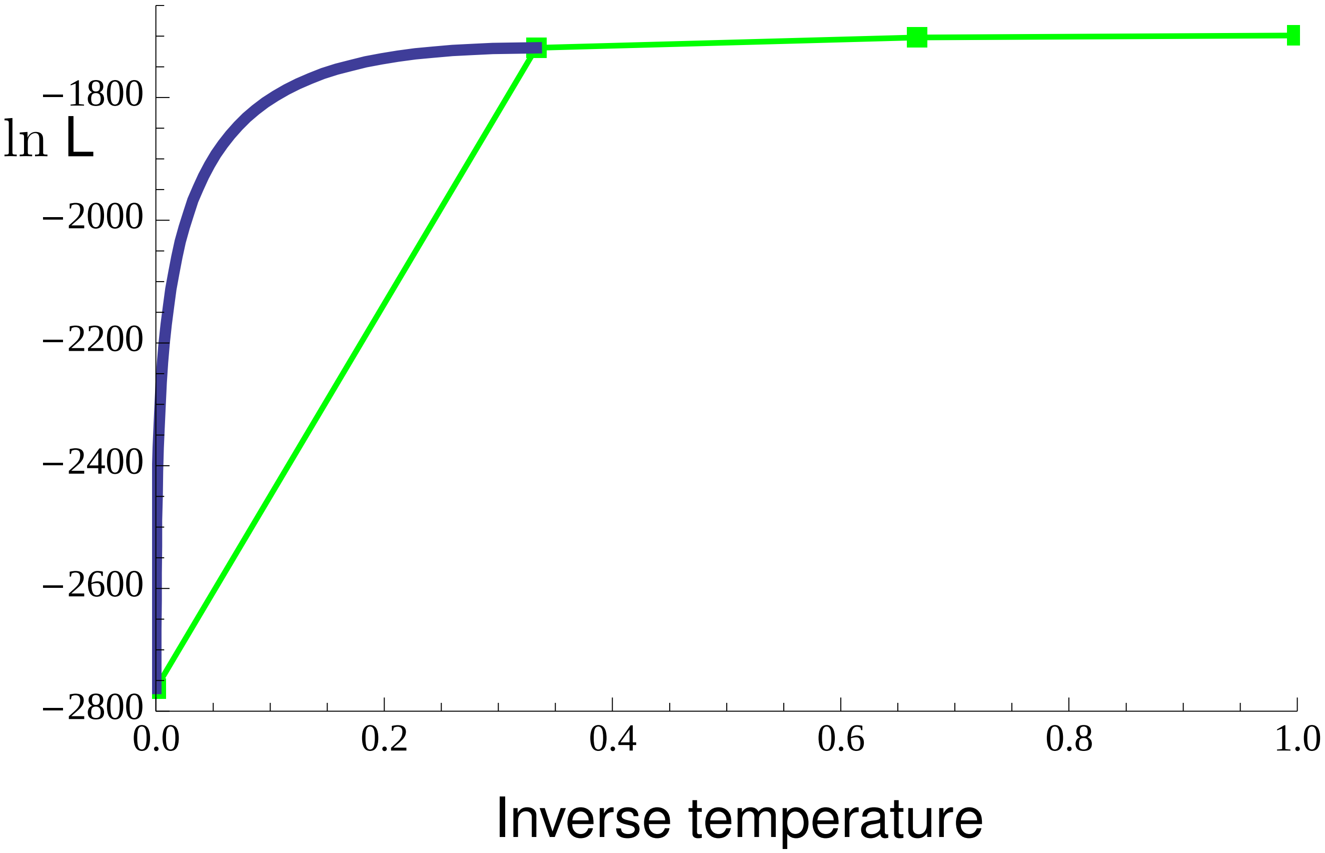
Thermodynamic integration



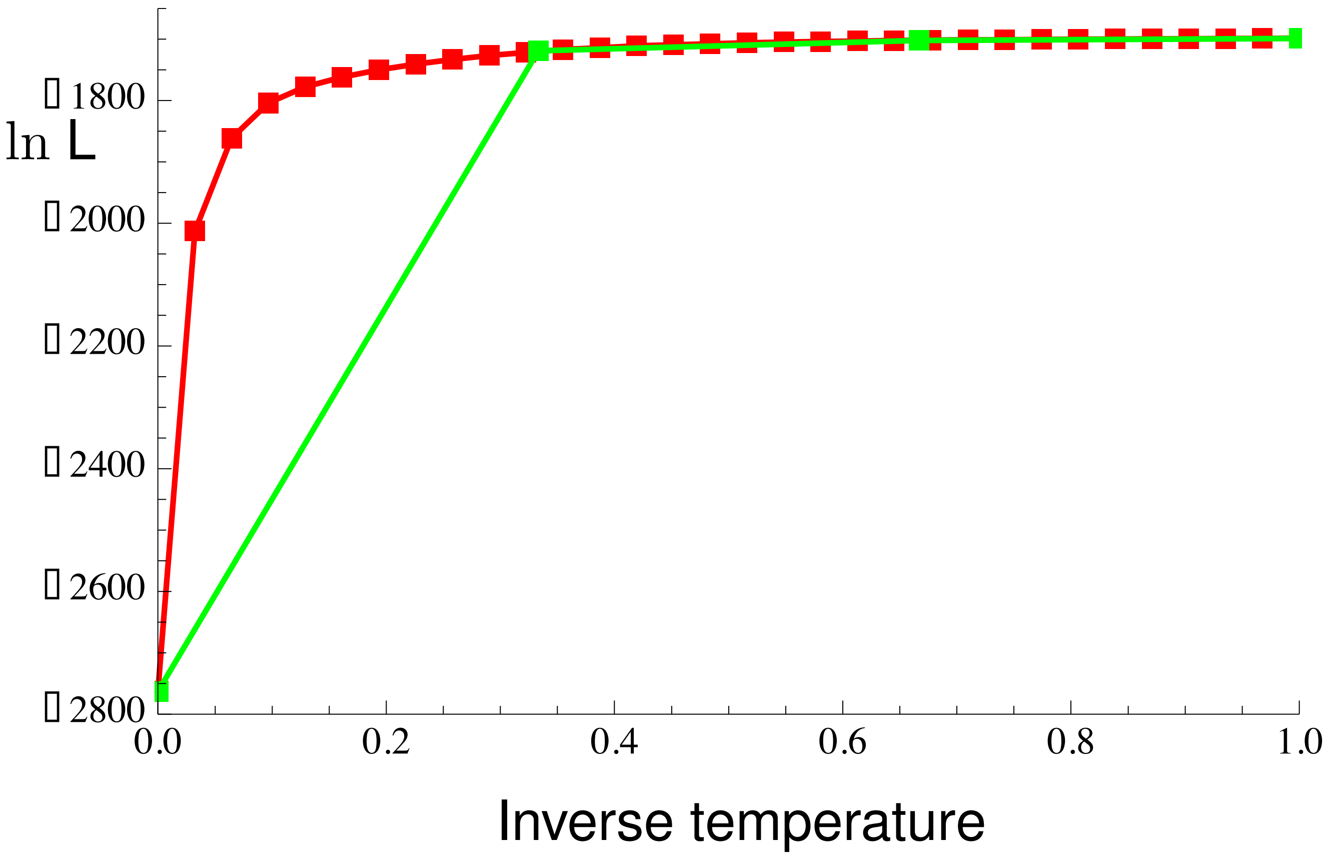
Thermodynamic integration



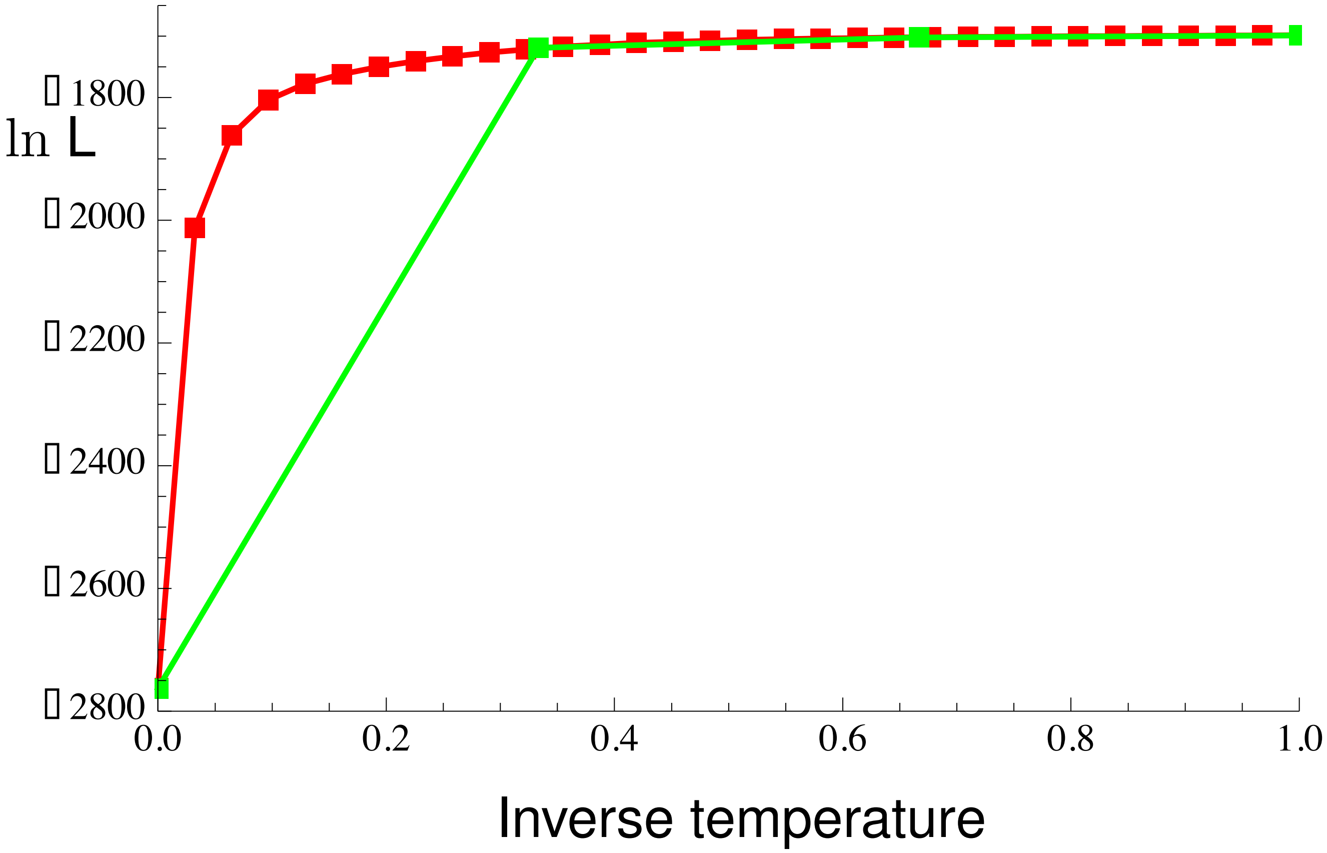
Thermodynamic integration



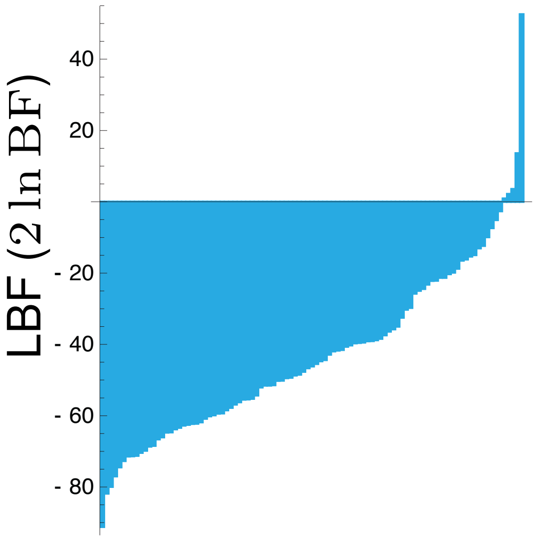
Thermodynamic integration



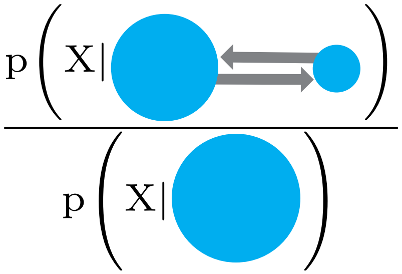
Thermodynamic integration



Bayes factor

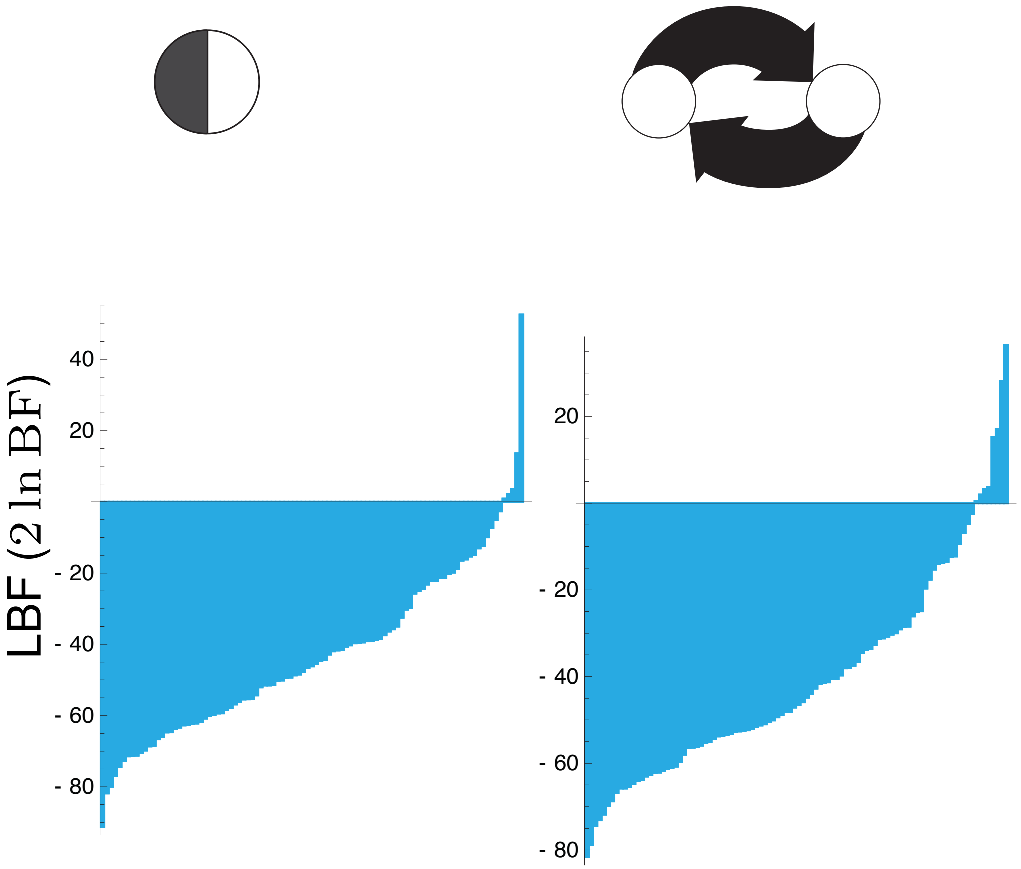


$$\text{LBF} = 2\ln\frac{p(X|M_1)}{p(X|M_2)} = 2\ln$$

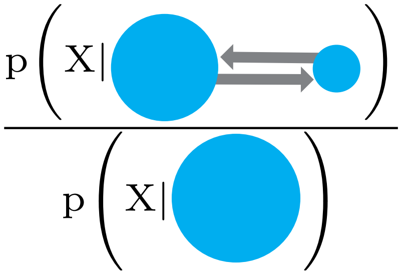


Bayes factor

Simulation results

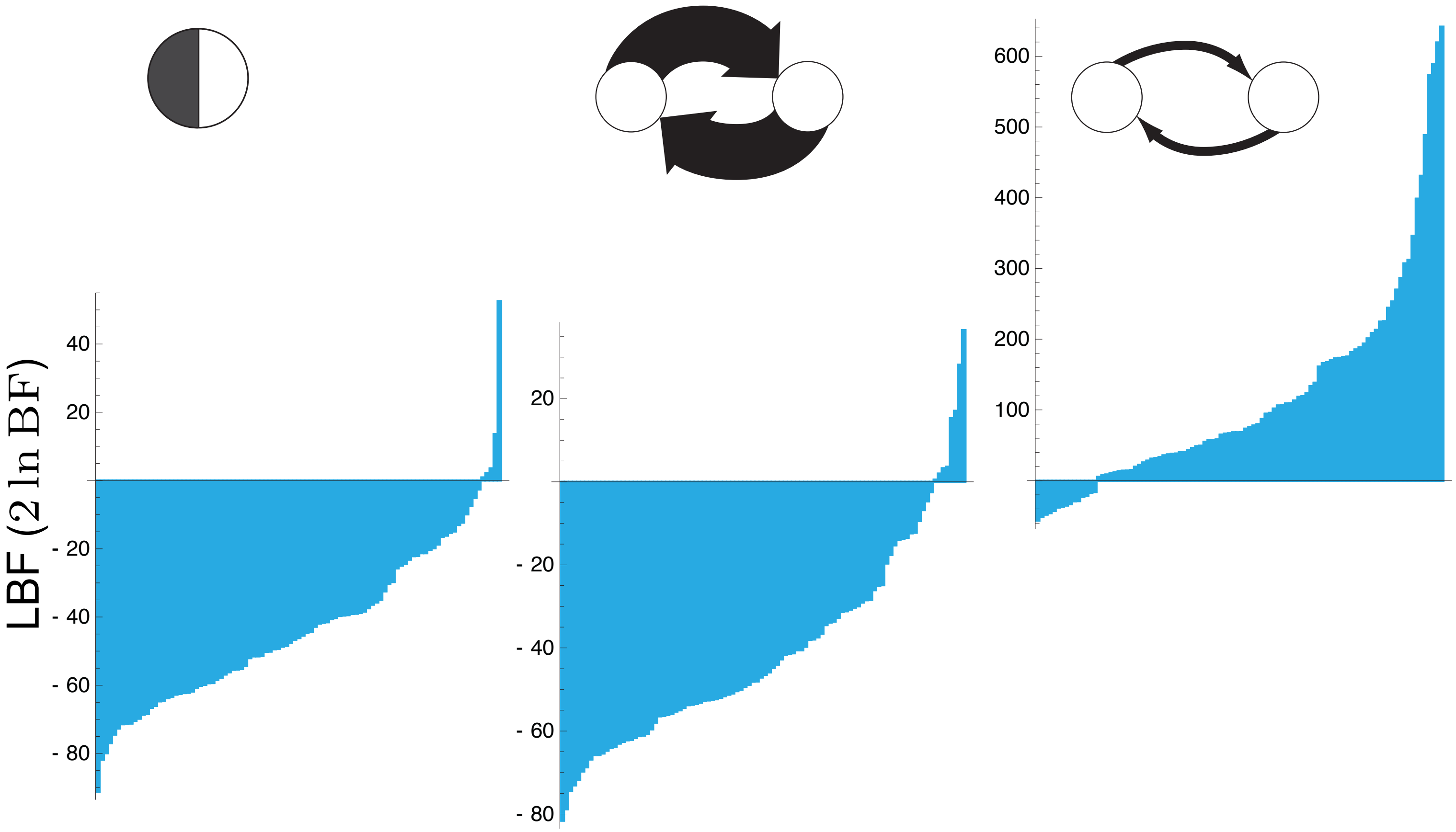


$$\text{LBF} = 2\ln\frac{p(X|M_1)}{p(X|M_2)} = 2\ln$$

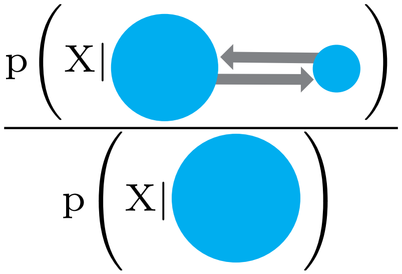


Bayes factor

Simulation results

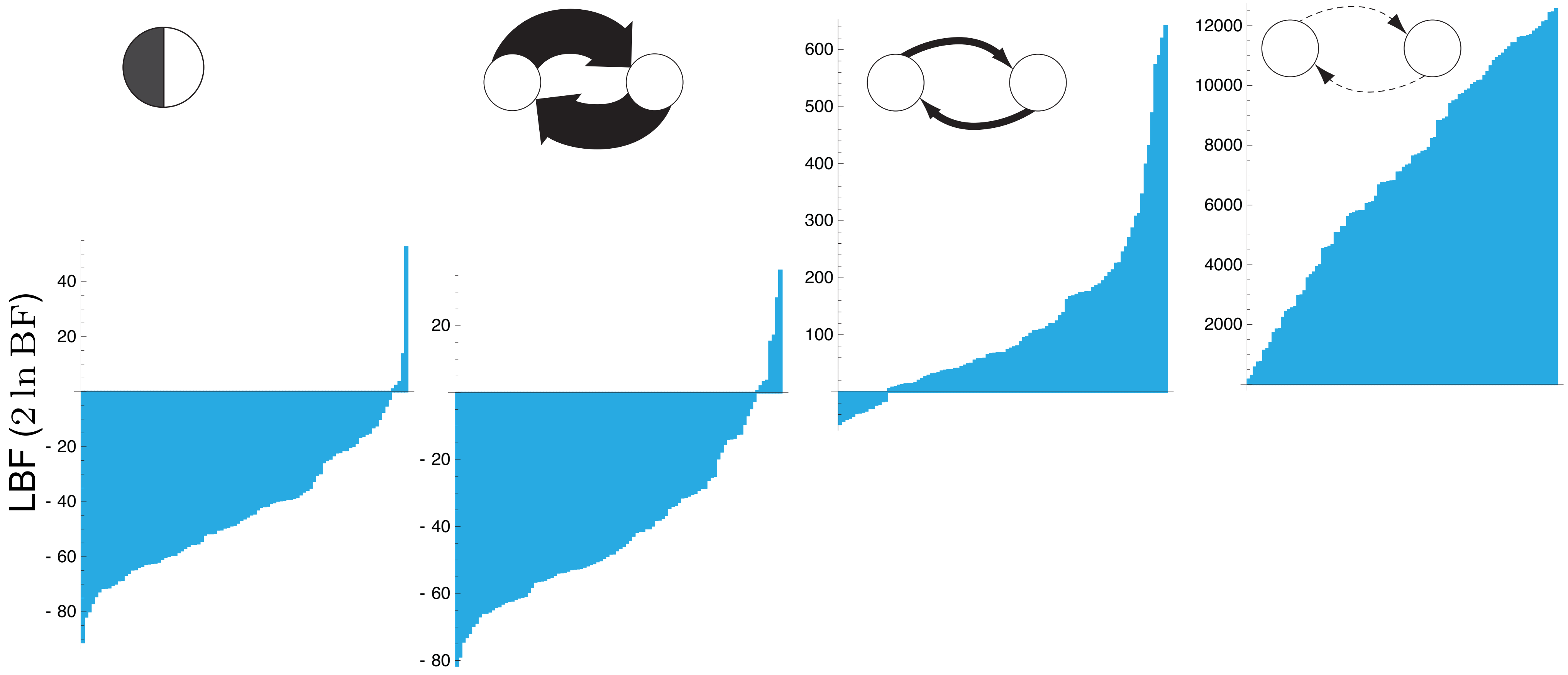


$$\text{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$

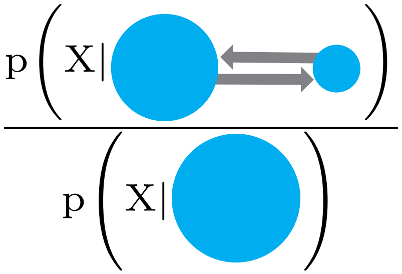


Bayes factor

Simulation results



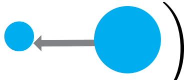
$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$

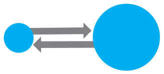
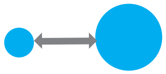
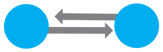
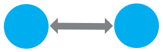
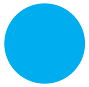
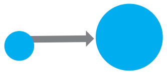
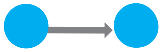
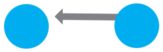


Bayes factor

Percent of Models

$$\left[\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}\left(\mathrm{X}|\qquad\right)}\right]$$



Param.	4	3	3	2	1	3	2	2
Model								
	xxxx	xmmx	mxxm	mmmm	x	x0xx	m0xm	mx0m
Rejected	100	100	100	100	97	71	46	29
Accepted	0	0	0	0	3	29	54	71

Total 20 sequences with length of 1000 bp

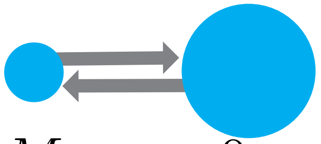
Parameters used to generate data:

$\Theta_i = 4N_e^{(i)}\mu$; $M_{ji} = \frac{m_{ji}}{\mu}$;

$Nm = \Theta M/4$

$M_{1\to 2} = 100$

$\Theta_1 = 0.005$

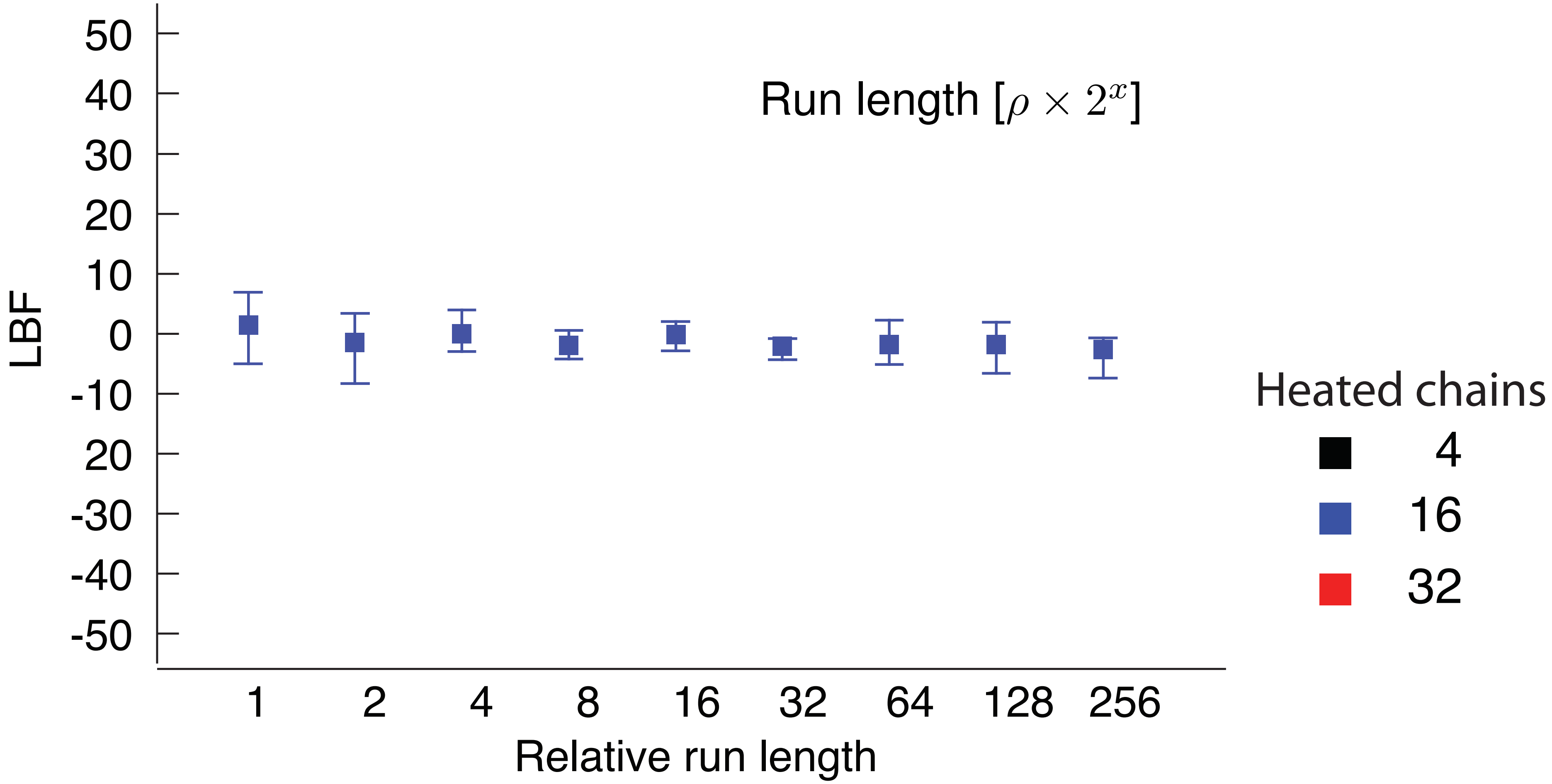


$\Theta_1 = 0.01$

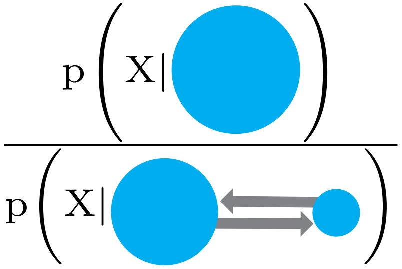
$M_{2\to 1} = 0$

Bayes factor: influence of runlength

Harmonic



$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$

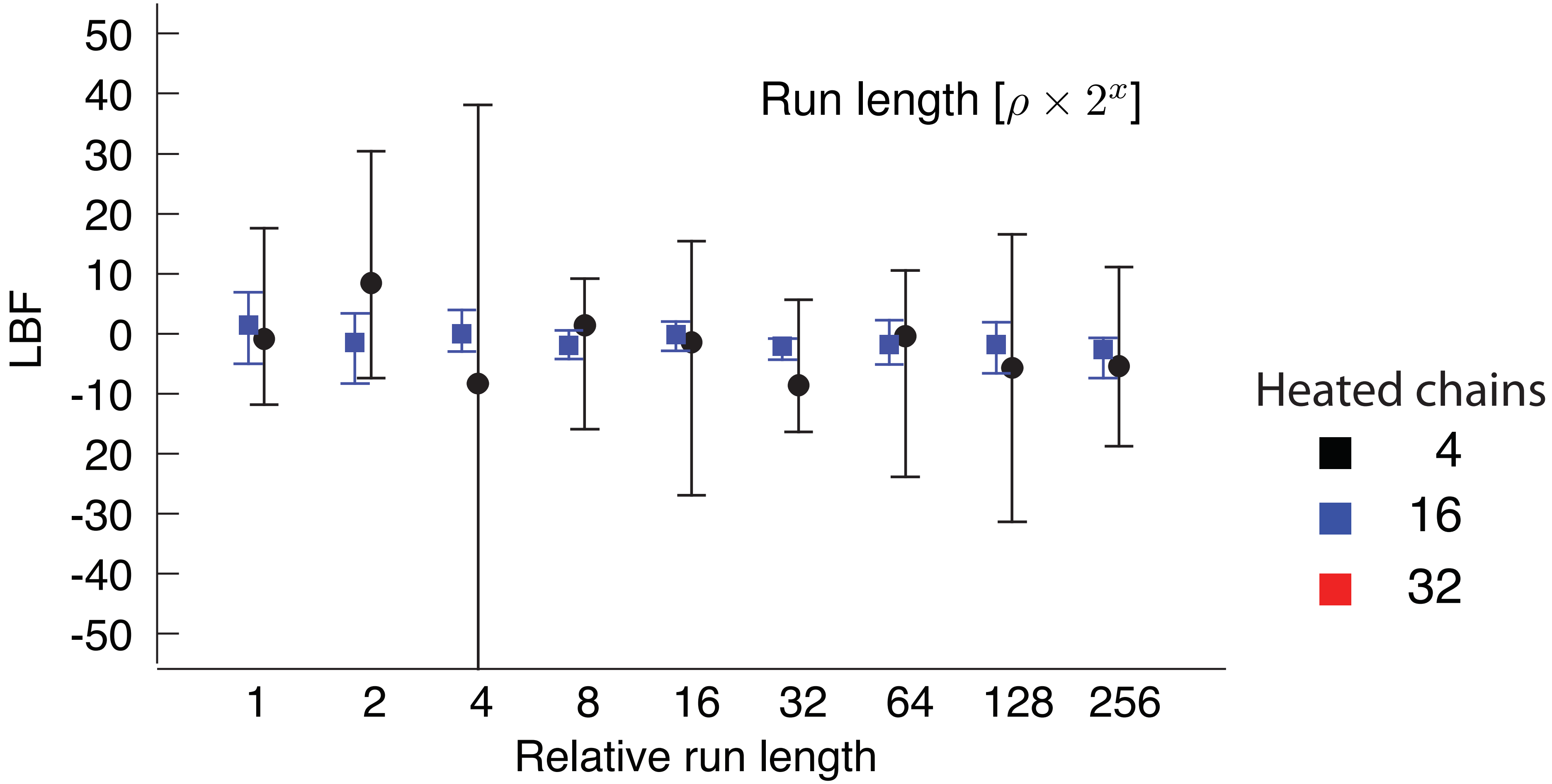


$\rho = (10^4 + 2 \times 10^4)$

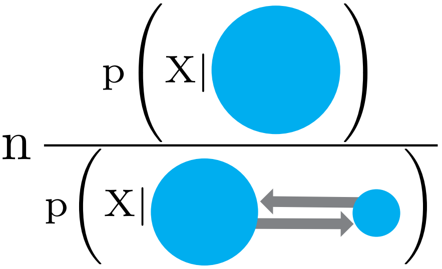
Time: 17 to 350 sec

Bayes factor: influence of runlength

Harmonic



$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$

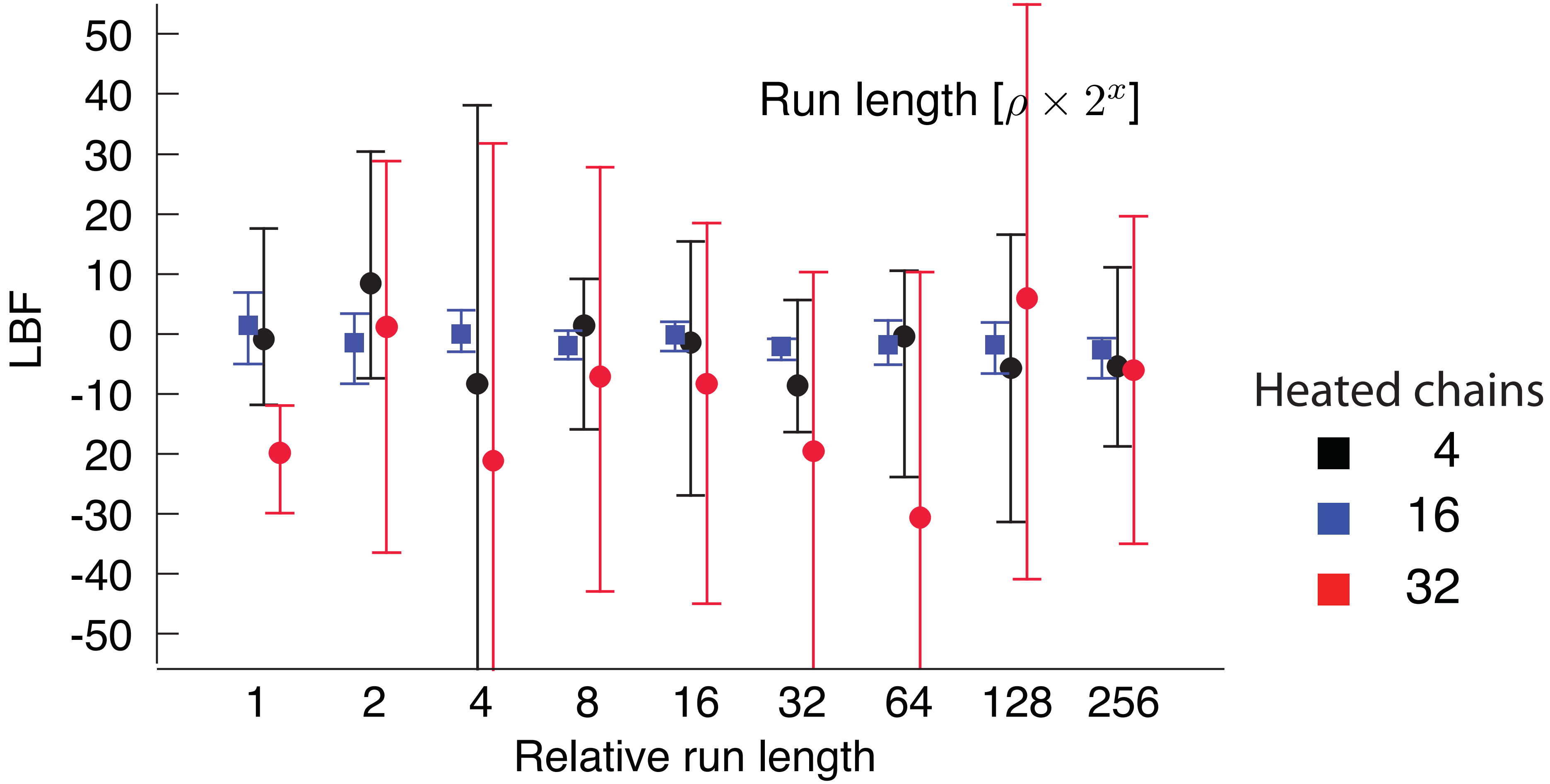


$\rho = (10^4 + 2 \times 10^4)$

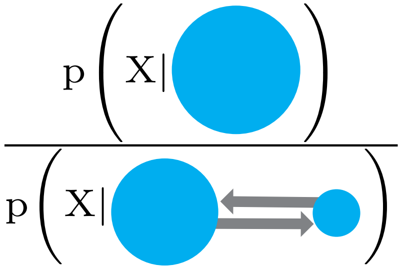
Time: 17 to 350 sec

Bayes factor: influence of runlength

Harmonic



$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$

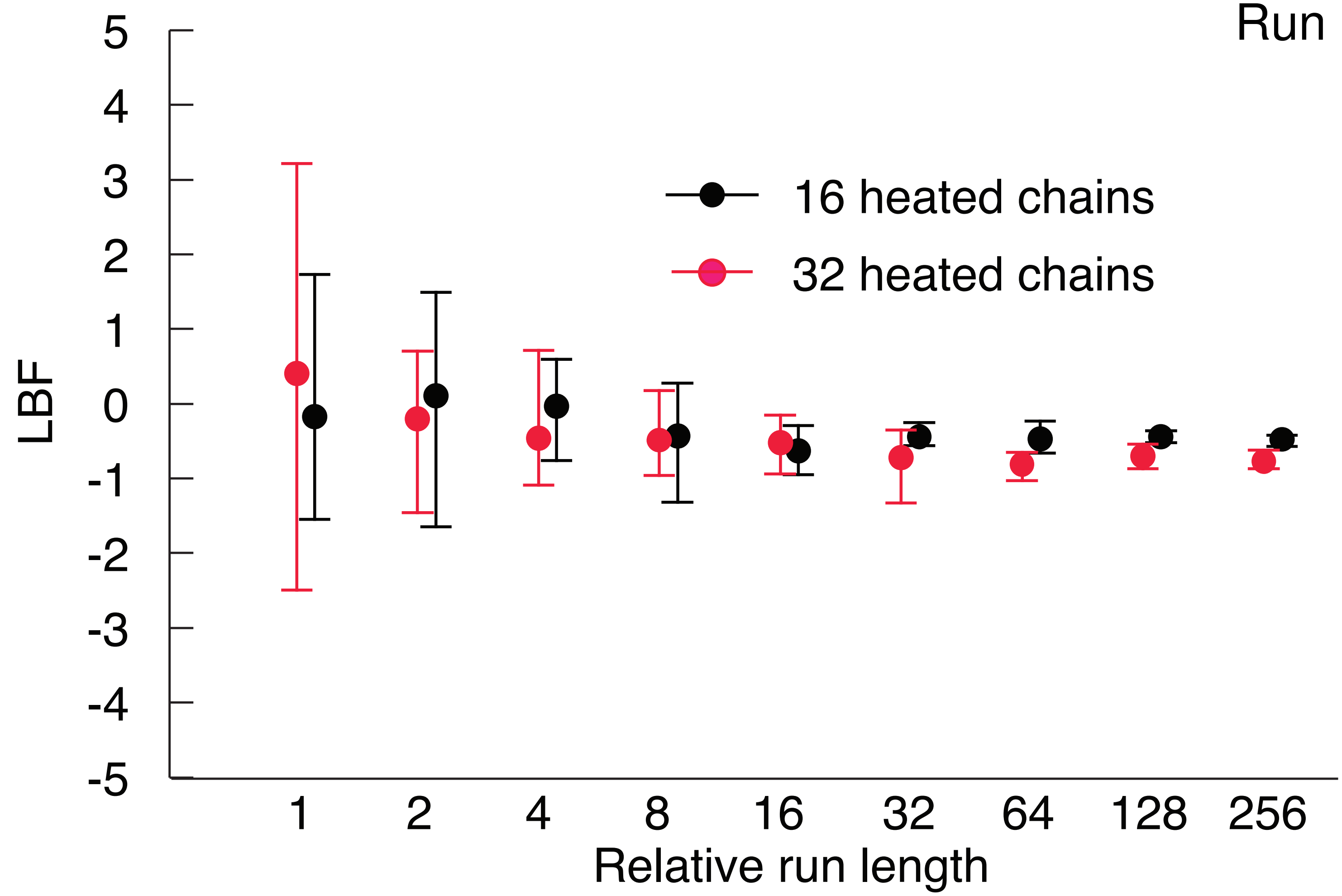


$\rho = (10^4 + 2 \times 10^4)$

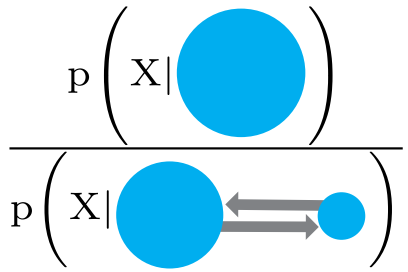
Time: 17 to 350 sec

Bayes factor: influence of runlength

Run length $[\rho \times 2^x]$



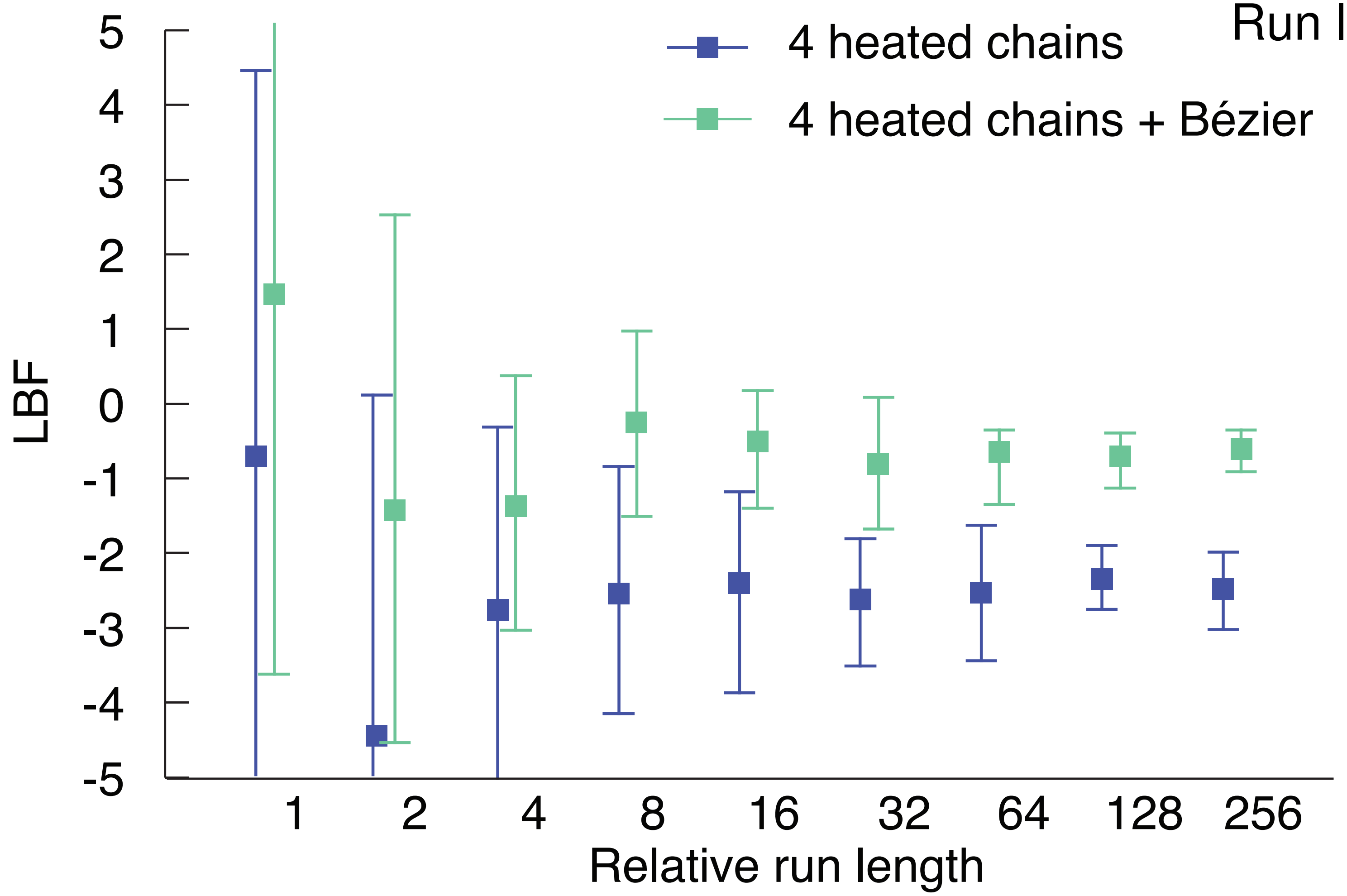
$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$



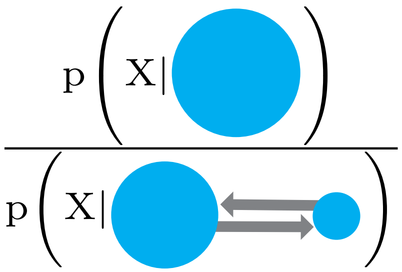
$$\rho = (10^4 + 2 \times 10^4)$$

Bayes factor: influence of runlength

Run length $[\rho \times 2^x]$



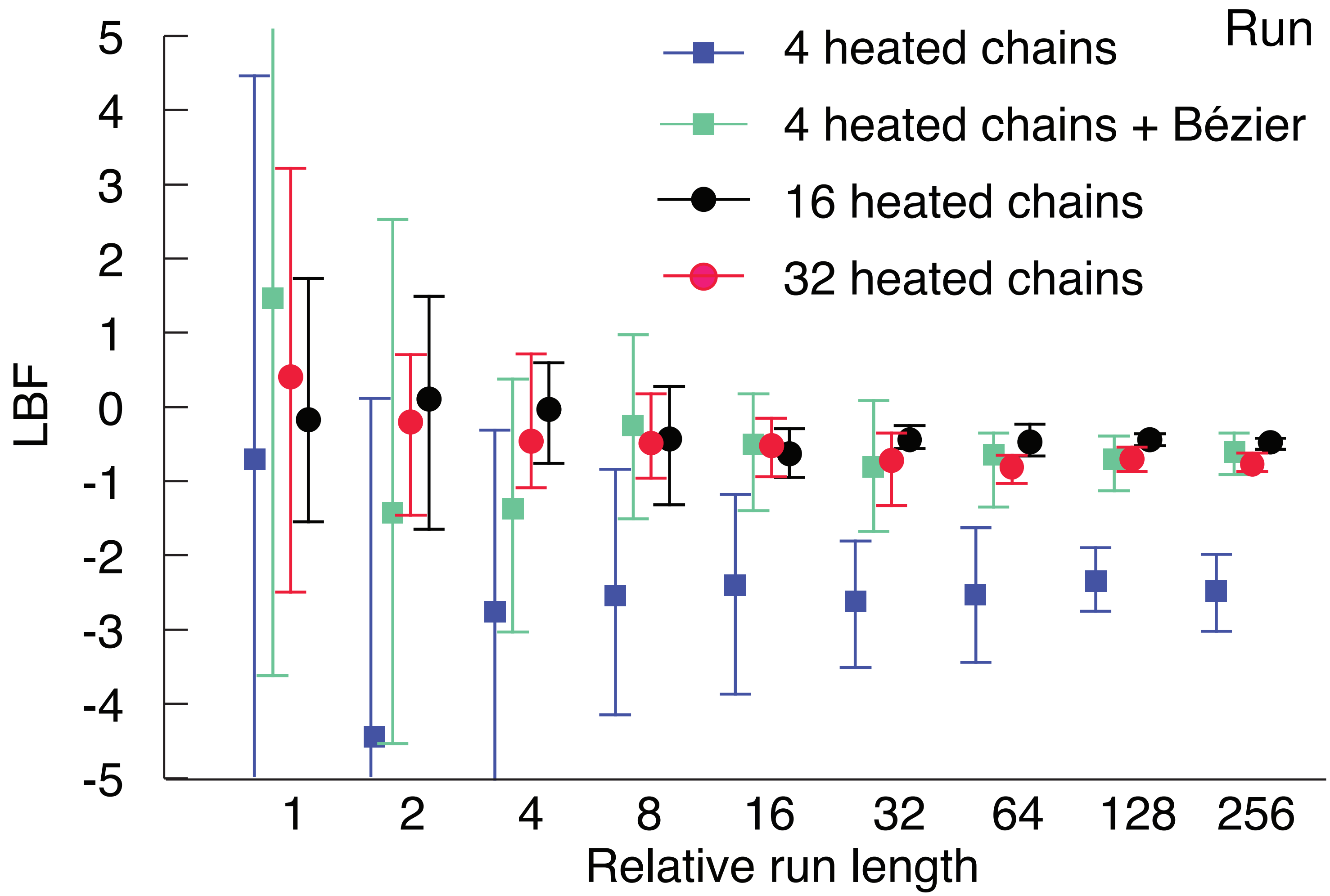
$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$



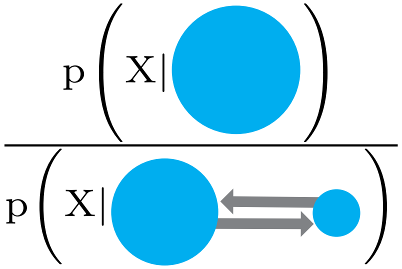
$$\rho = (10^4 + 2 \times 10^4)$$

Bayes factor: influence of runlength

Run length $[\rho \times 2^x]$

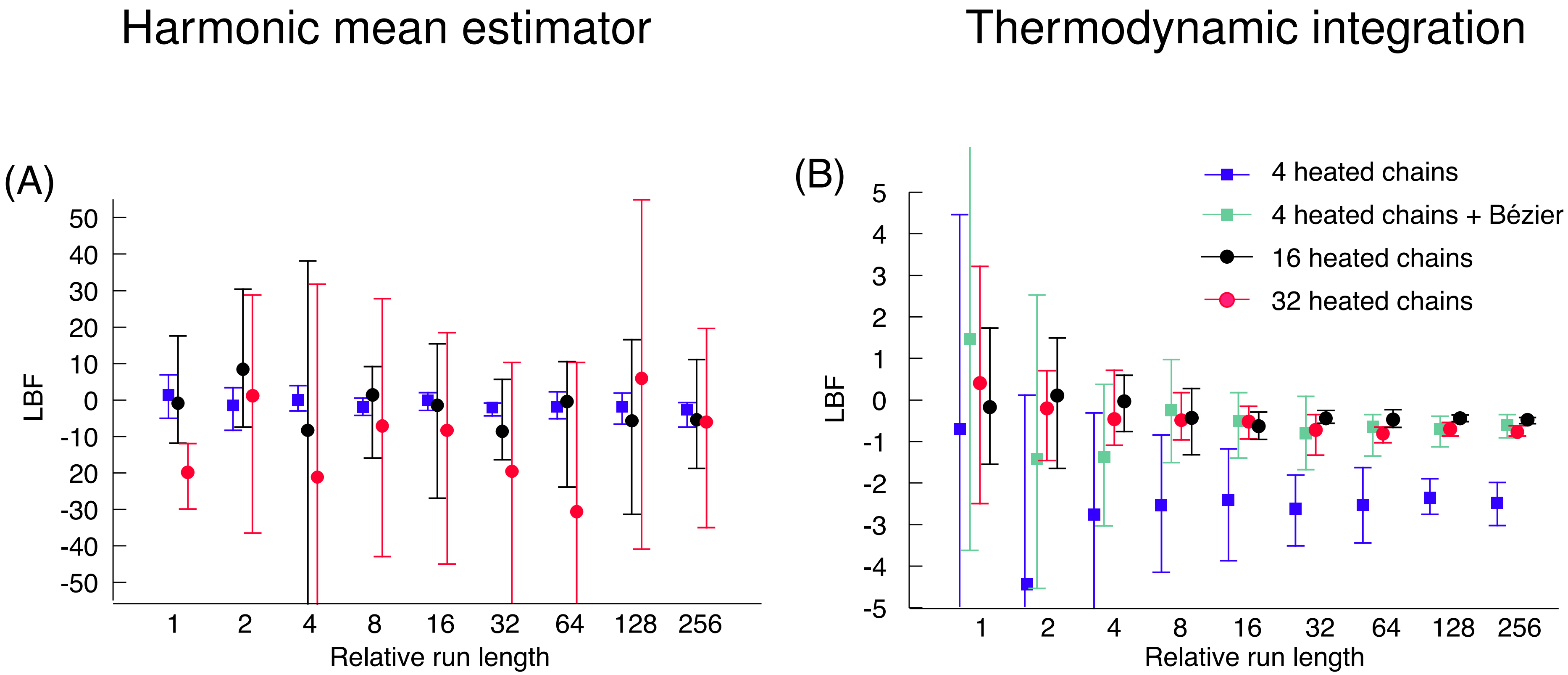


$$\mathrm{LBF} = 2\ln\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} = 2\ln$$



$$\rho = (10^4 + 2 \times 10^4)$$

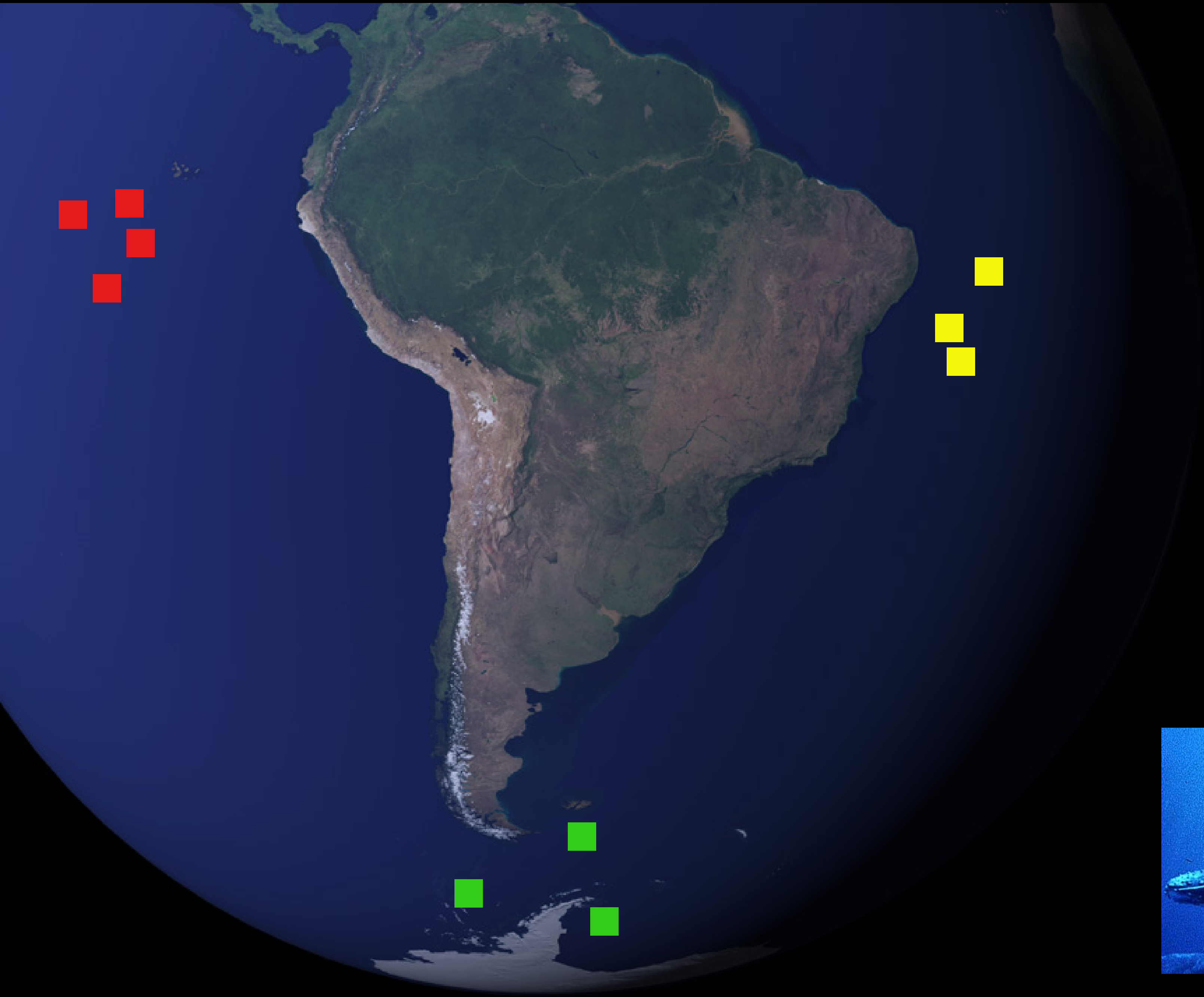
Bayes factor: influence of runlength


Harmonic mean estimator

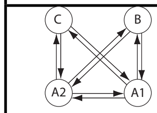
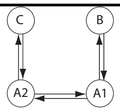
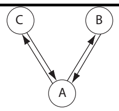
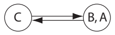
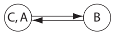
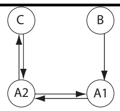
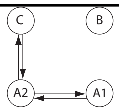
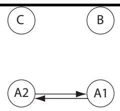
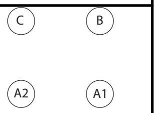
Thermodynamic integration

Humpback whales in the South Atlantic

Real data



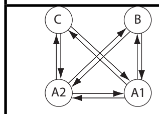
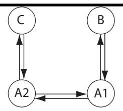
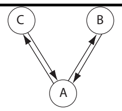
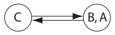
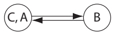
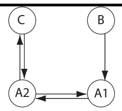
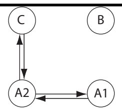
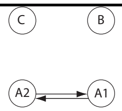
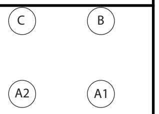
Using Marginal Likelihoods to rank

Replica[1]	$\hat{\ell}_{M_i}$ of models M_i								
									
1 (10)									
1’ (10)[2]									
2 (10)									
3 (30)									
Rank									

[1] Number of samples per population in parentheses.

[2] Same data as in replicate 1, but different start values of MCMC run.

Using Marginal Likelihoods to rank

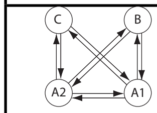
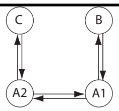
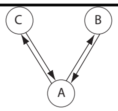
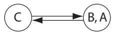
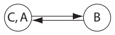
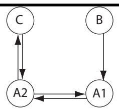
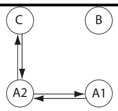
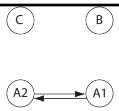
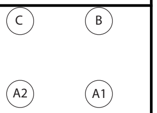
Replica[1]	$\hat{\ell}_{M_i}$ of models M_i								
									
1 (10)	-1988	-1958	-1984	-2009	-2054	-1935	-2070	**-1793**	-2015
1' (10)[2]									
2 (10)									
3 (30)									
Rank	5	3	4	6	8	2	9	1	7

[1] Number of samples per population in parentheses.

[2] Same data as in replicate 1, but different start values of MCMC run.

Using Marginal Likelihoods to rank

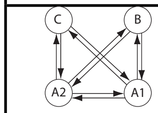
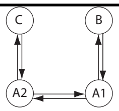
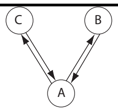
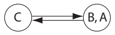
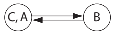
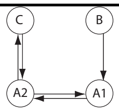
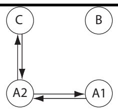
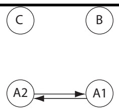
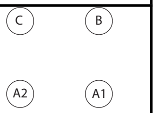
Whale example

Replica[1]	$\hat{\ell}_{M_i}$ of models M_i								
									
1 (10)	-1988	-1958	-1984	-2009	-2054	-1935	-2070	**-1793**	-2015
1’ (10)[2]	-1988	-1958	-1984	-2009	-2054	-1936	-2070	**-1793**	-2002
2 (10)									
3 (30)									
Rank	5	3	4	6	8	2	9	1	7

[1] Number of samples per population in parentheses.

[2] Same data as in replicate 1, but different start values of MCMC run.

Using Marginal Likelihoods to rank

Replica[1]	$\hat{\ell}_{M_i}$ of models M_i								
									
1 (10)	-1988	-1958	-1984	-2009	-2054	-1935	-2070	**-1793**	-2015
1' (10)[2]	-1988	-1958	-1984	-2009	-2054	-1936	-2070	**-1793**	-2002
2 (10)	-2034	-2005	-2030	-2056	-2099	-1985	-2134	**-1856**	-2071
3 (30)	-3669	-3519	-3630	-3735	-3983	-3454	-3689	**-2725**	-3028
Rank	7	5	6	8	9	3	4	1	2

[1] Number of samples per population in parentheses.

[2] Same data as in replicate 1, but different start values of MCMC run.