Bayesian Model Selection

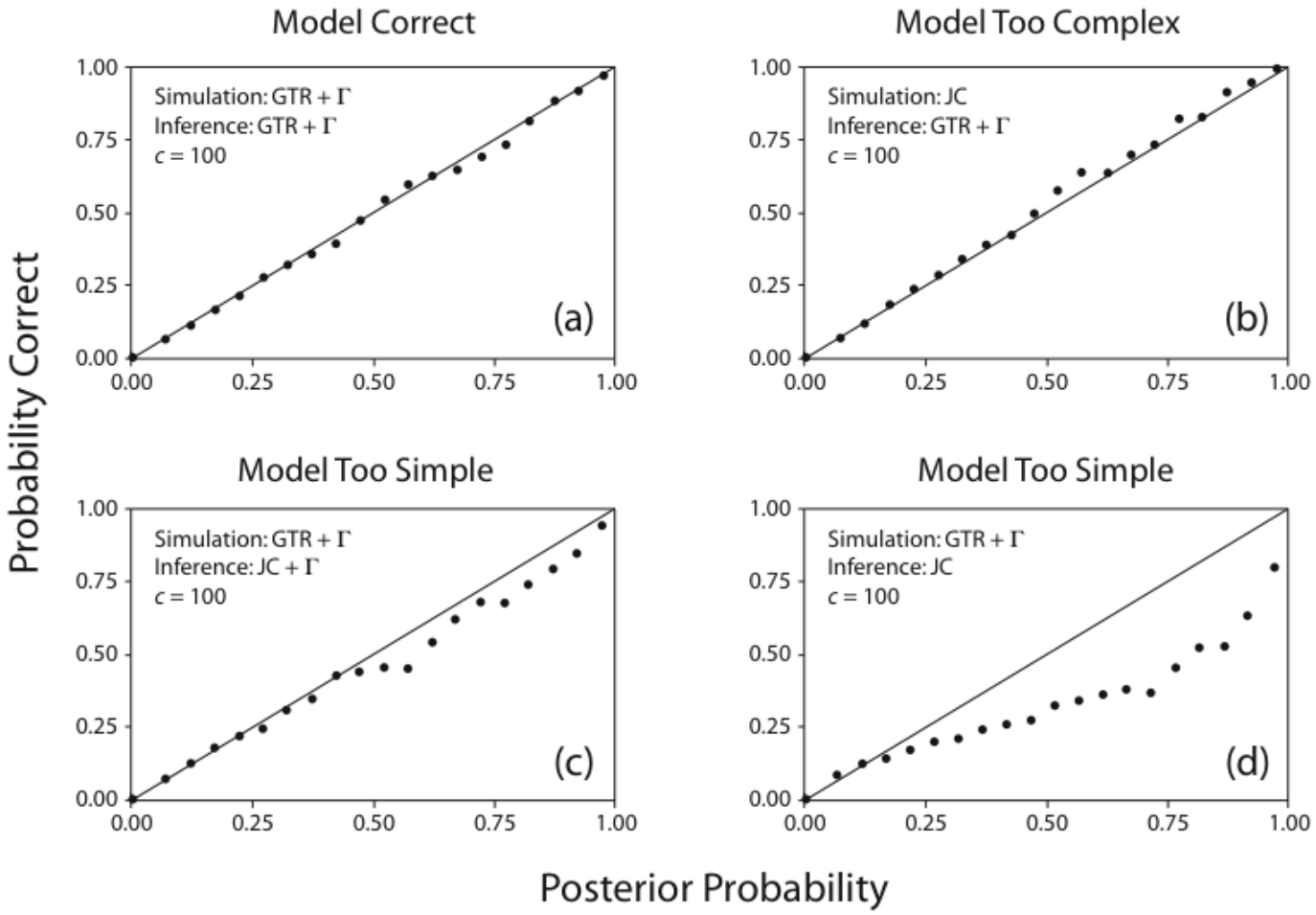
Fred Ronquist
Swedish Museum of Natural History,
Stockholm, Sweden

Workshop in Advanced Bayesian Phylogenetics
Adelaide, Nov. 17-21, 2014

Topics

- Bayesian model comparison
- The model as a random variable
- Model adequacy

Bayesian Model Sensitivity



Models, models, models

- Alignment-free models
- Heterogeneity in substitution rates and stationary frequencies across sites and lineages
- Relaxed clock models
- Models for morphology and biogeography
- Models describing model spaces, e.g. GTR space and partition space
- Models of dependence across sites according to 3D structure of proteins
- Positive selection models
- Amino acid models
- Models for population genetics and phylogeography

1. Bayesian Model Comparison

Model testing with ML

- Hierarchical likelihood ratio test compares maximum likelihood (L) across models (models need to be nested)

$$2 \ln L_i$$

- Akaike Information Criterion (AIC):

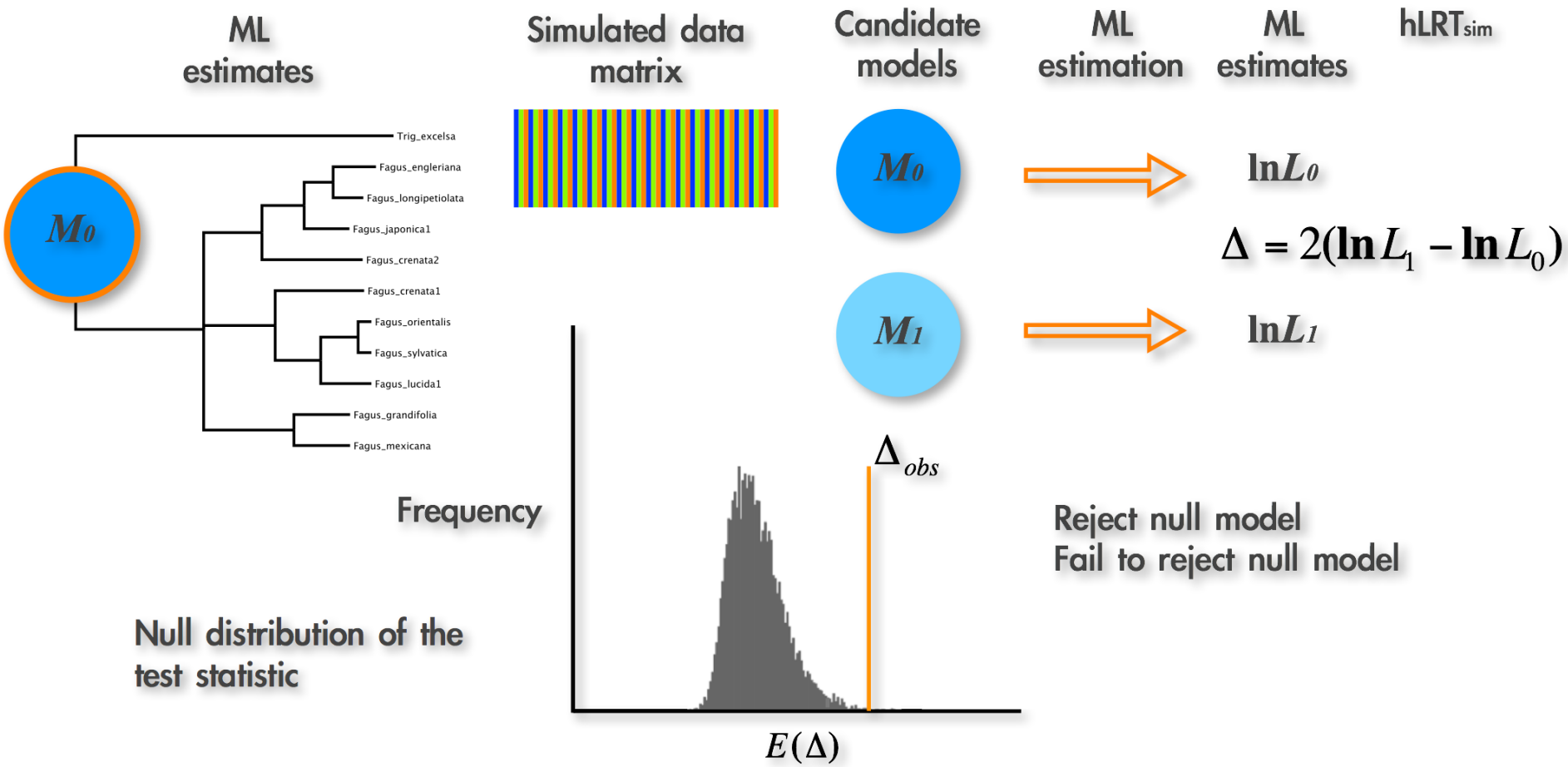
$$AIC_i = -2 \ln L_i + 2 p_i$$

- Bayesian Information Criterion (BIC):

$$BIC_i = -2 \ln L_i + p_i \ln n_i$$

- AIC and BIC penalize the maximum likelihood (*L*) with the number of parameters (*p*) and the amount of data (*n*) -> balance between overfitting and error variance

Parametric Bootstrapping



Bayes' Rule

$$f(\theta \mid D) = \frac{f(\theta) f(D \mid \theta)}{\int f(\theta) f(D \mid \theta)\, \mathrm{d}\theta} = \frac{f(\theta) f(D \mid \theta)}{f(D)}$$

Marginal likelihood (of the data)

We have implicitly conditioned on a model:

$$f(\theta \mid D, M) = \frac{f(\theta \mid M) f(D \mid \theta, M)}{f(D \mid M)}$$

$$\Pr(D \mid M) \quad \xrightarrow{?} \quad \Pr(M \mid D)$$

Bayes' rule:

$$\Pr(M \mid D) = \frac{\Pr(M)\Pr(D \mid M)}{\sum \Pr(M)\Pr(D \mid M)}$$

Bayesian Model Choice

Posterior model odds:

$$\frac{\Pr(M_1 \mid D)}{\Pr(M_0 \mid D)} = \frac{\Pr(M_1)\Pr(D \mid M_1)}{\Pr(M_0)\Pr(D \mid M_0)}$$

Bayes factor:

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

Bayesian Model Choice

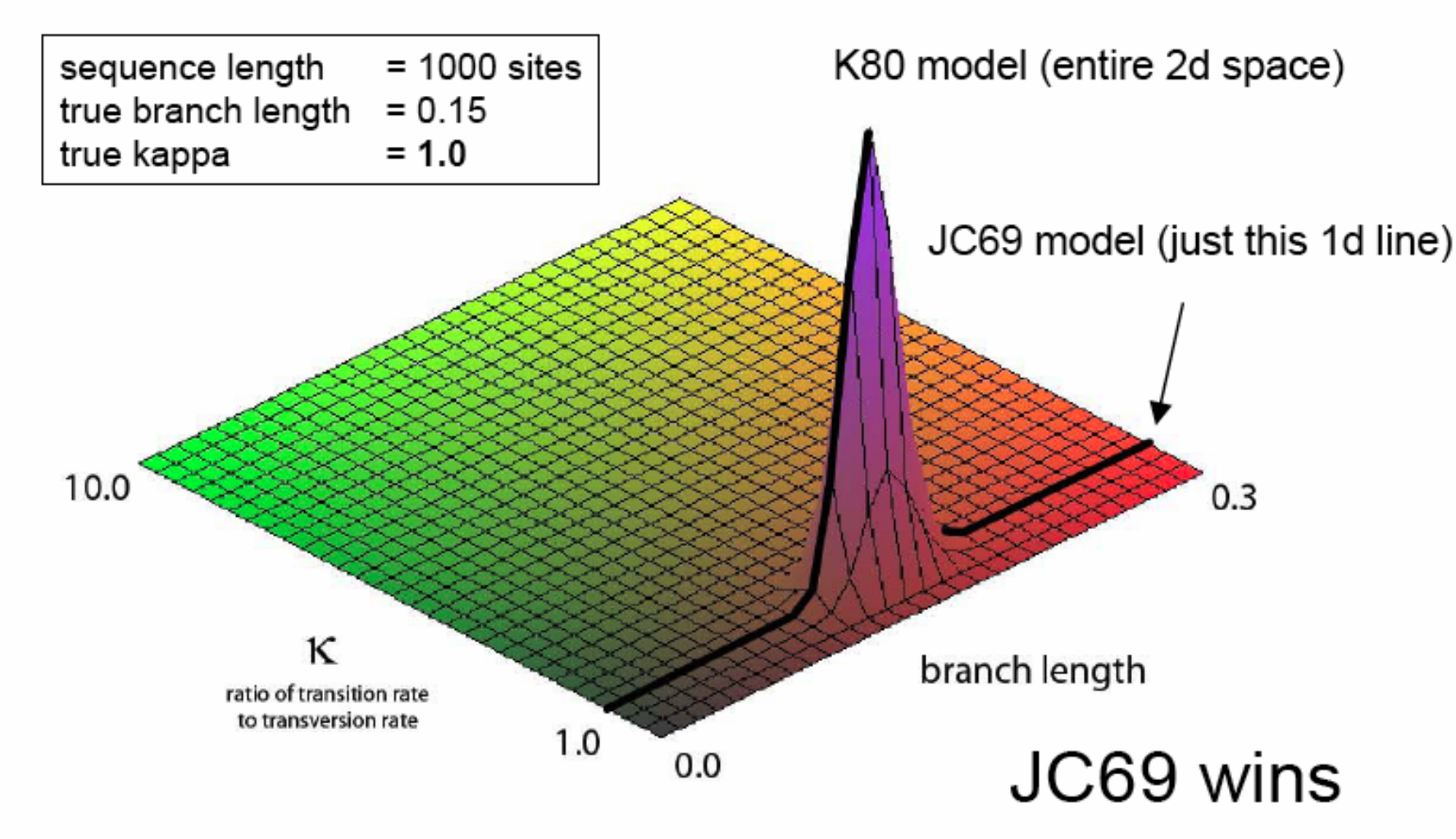
- The normalizing constant in Bayes' theorem, the marginal likelihood of the data, Pr(D) or Pr(D|M), can be used for model choice
- What we are comparing is the average likelihood of the data across the prior
- Any models can be compared: nested, non-nested, data-derived; it is just a probability comparison
- No correction for number of parameters
- Can prefer a simpler model over a more complex model (comparing average likelihoods)
- Critical values in Kass and Raftery (1997)

Bayes Factor Comparisons

Interpretation of the Bayes factor

$2\ln(B_{10})$	B_{10}	Evidence against M_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

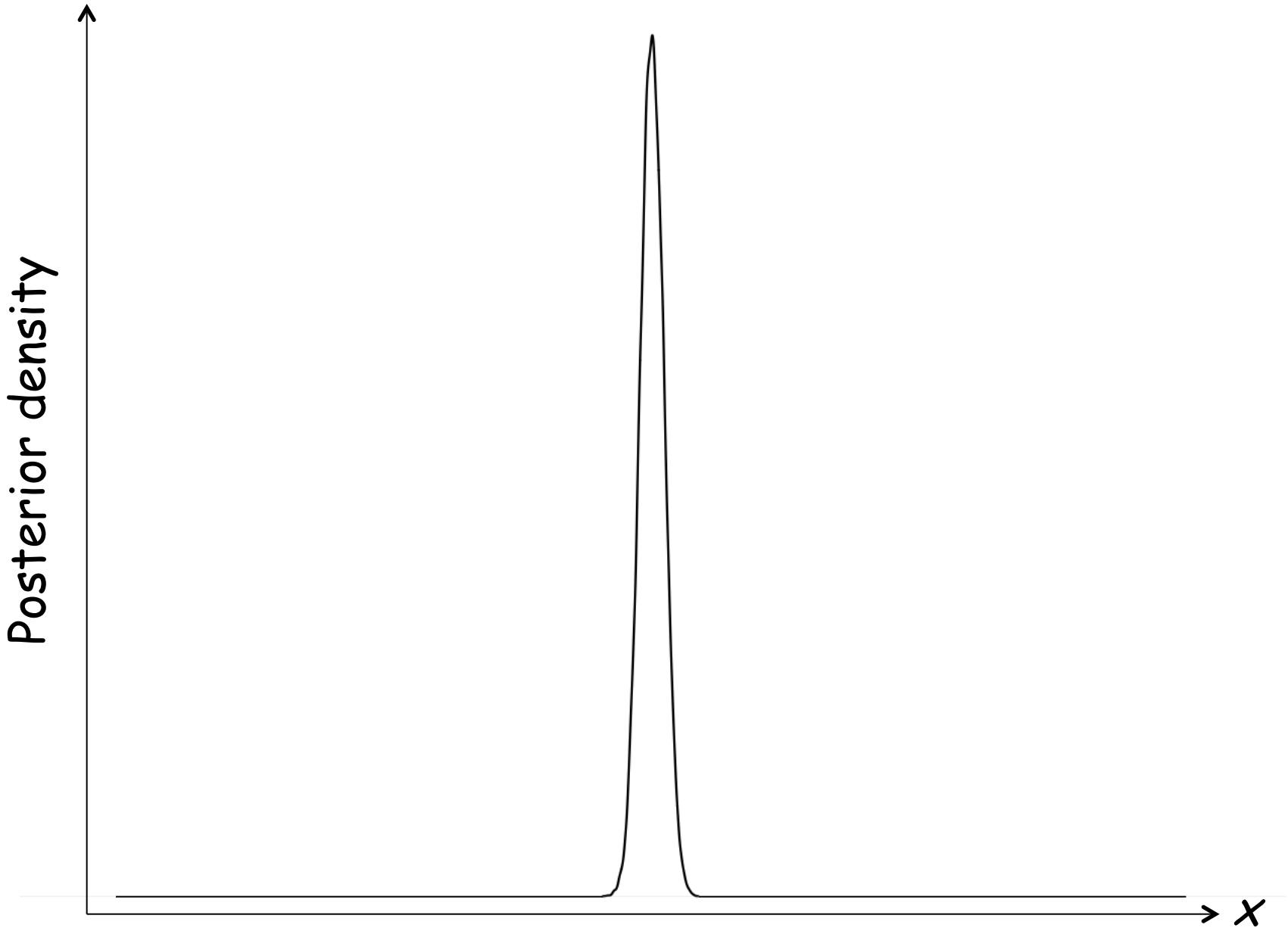
Simple Model Wins (from Lewis)

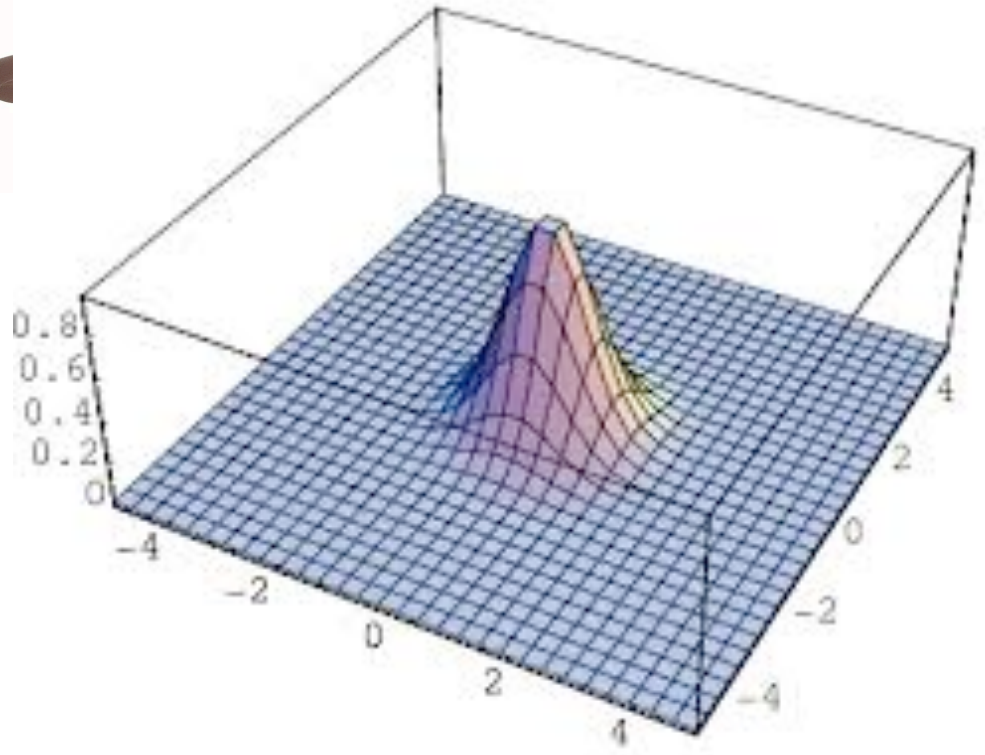


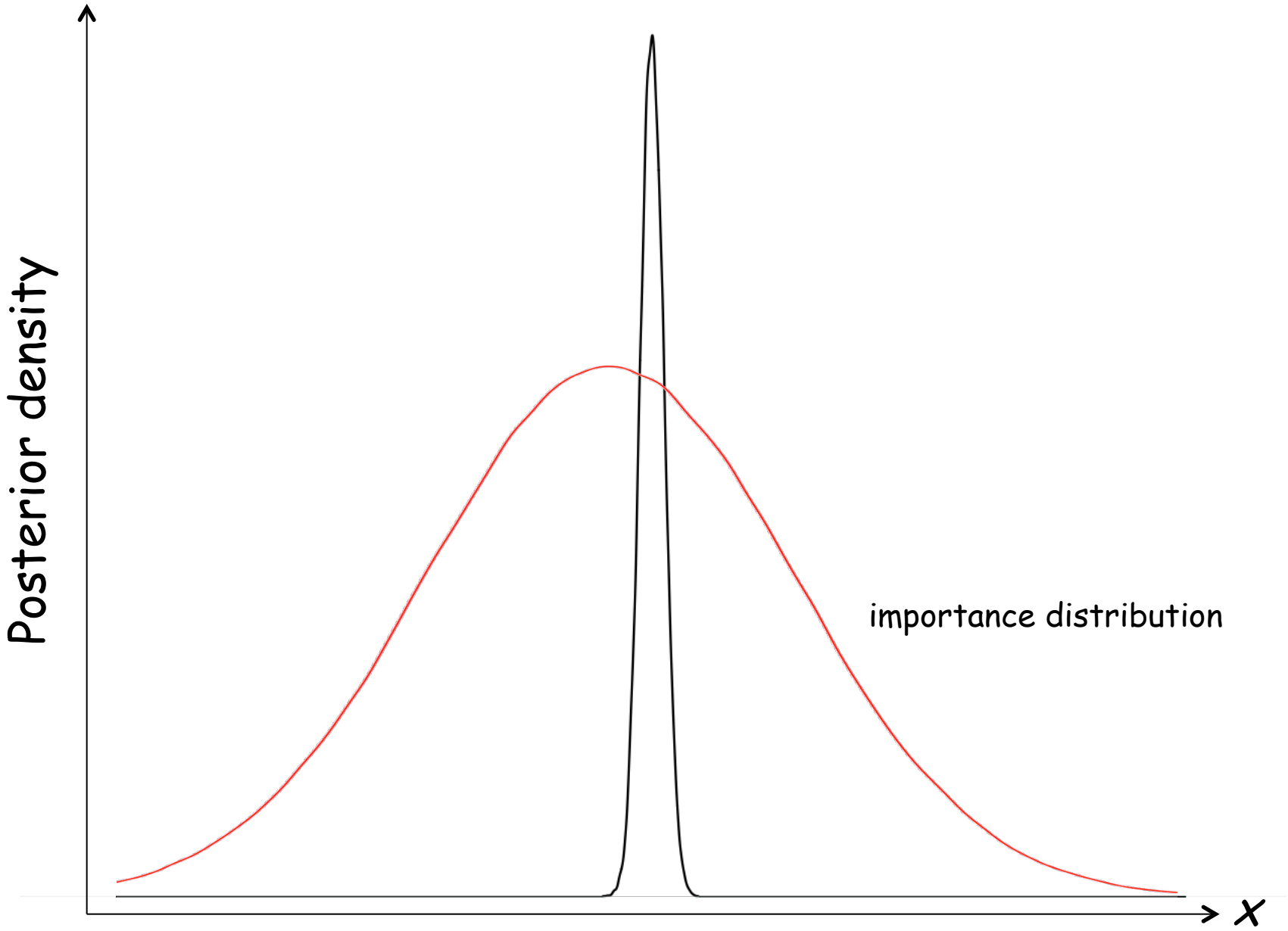

IF YOU VOTE FOR ME
ALL OF YOUR WILDEST
DREAMS WILL COME TRUE

Estimating Model Likelihood

- Harmonic mean estimator
- Arithmetic mean estimator
- Thermodynamic integration (path sampling)
- Stepping-stone sampling
- All except thermodynamic integration are based on importance sampling


Posterior density
X


0.8
0.6
0.4
0.2
0
-4
-2
0
2
4
4
2
0
-2
-4


Posterior density
importance distribution
X

Approximation using Importance Sampling

$g(\theta)$ Importance distribution

$$f\left(D \mid M\right) = \frac{E_g\left[f(D \mid \theta, M) w(\theta \mid M)\right]}{E_g\left[w(\theta \mid M)\right]}$$

$$f\left(D \mid M\right) \approx \frac{\sum_i f(D \mid \theta_i, M) w_i(\theta_i \mid M)}{\sum_i w_i(\theta_i \mid M)}$$

Simple choices of importance sampling distribution

$g(\theta) = f(\theta \mid D, M)$ Importance distribution = posterior

$$\frac{1}{f\left(D \mid M\right)} \approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{f(D \mid \theta_i, M)}$$ Harmonic mean estimator

Unbiased in the limit but biased in practice

$g(\theta) = f(\theta \mid M)$ Importance distribution = prior

$$f\left(D \mid M\right) \approx \frac{1}{n}\sum_{i=1}^{n} f(D \mid \theta_i, M)$$ Arithmetic mean estimator

Unbiased but has unacceptably high variance

Stepping-Stone Sampling

$$q_\beta = f(D \mid \theta, M)^\beta f(\theta \mid M)$$

Power posterior density:

$$p_\beta = q_\beta / c_\beta \qquad c_\beta = \int f(D \mid \theta, M)^\beta f(\theta \mid M) \mathrm{d}\theta$$

Consider running an MCMC sampler on the prior, and then computing average of likelihoods

$$r_{ss} = \frac{c_{1.0}}{c_{0.0}}$$

$$c_{1.0} = f(D \mid M)$$

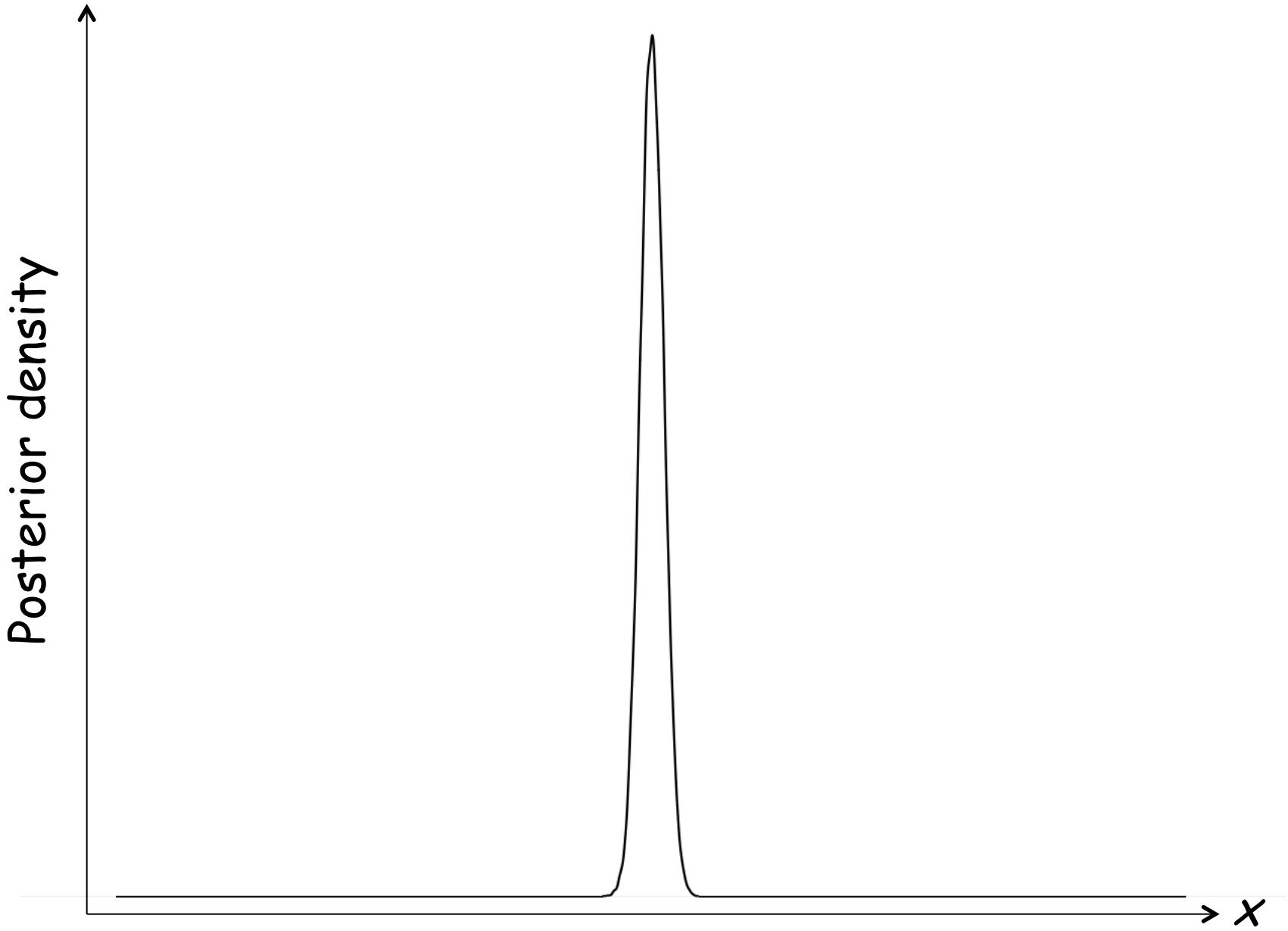
$$c_{0.0} = f(\theta \mid M) = 1.0$$

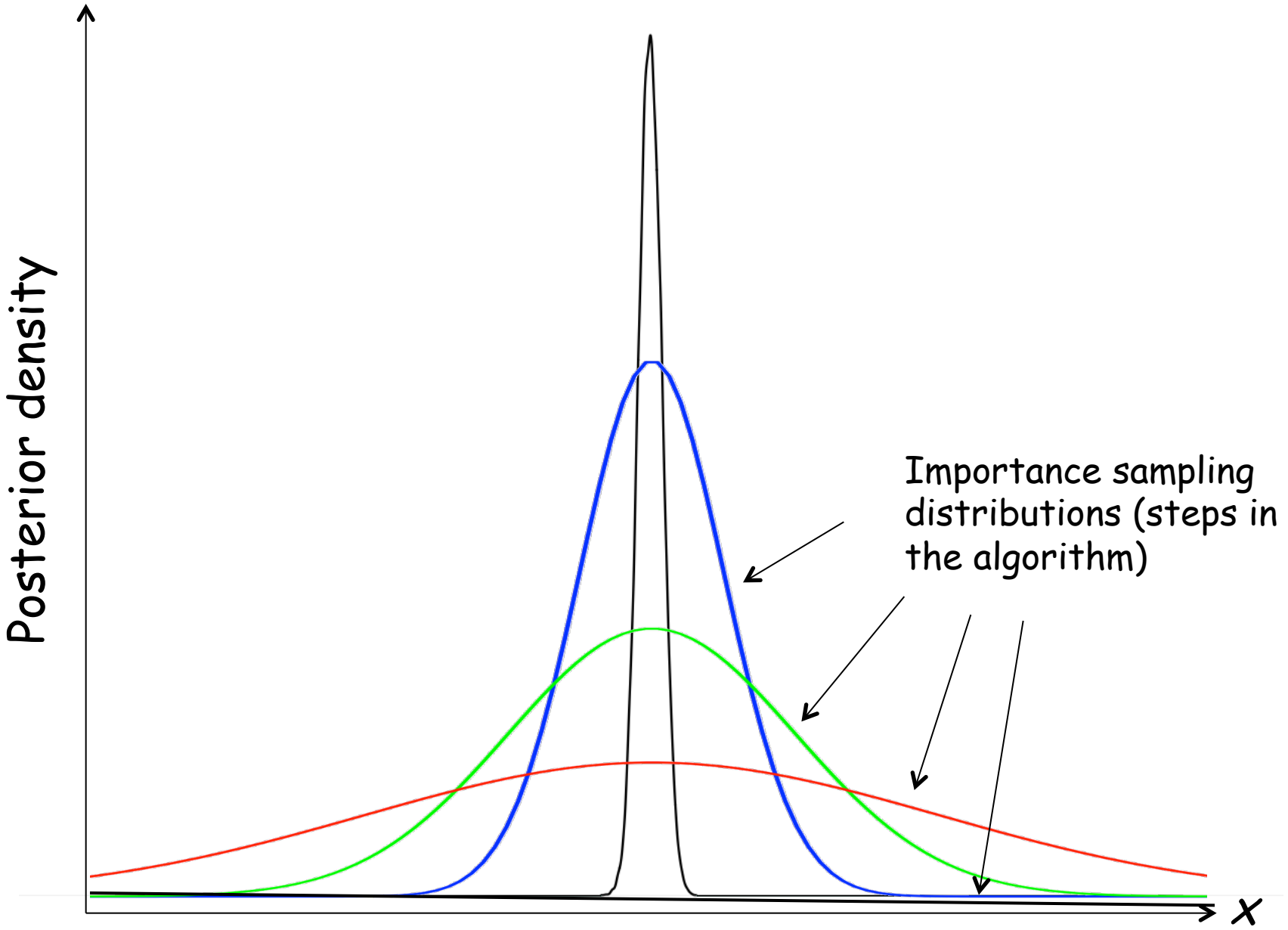
Single step sampling same as arithmetic mean estimator

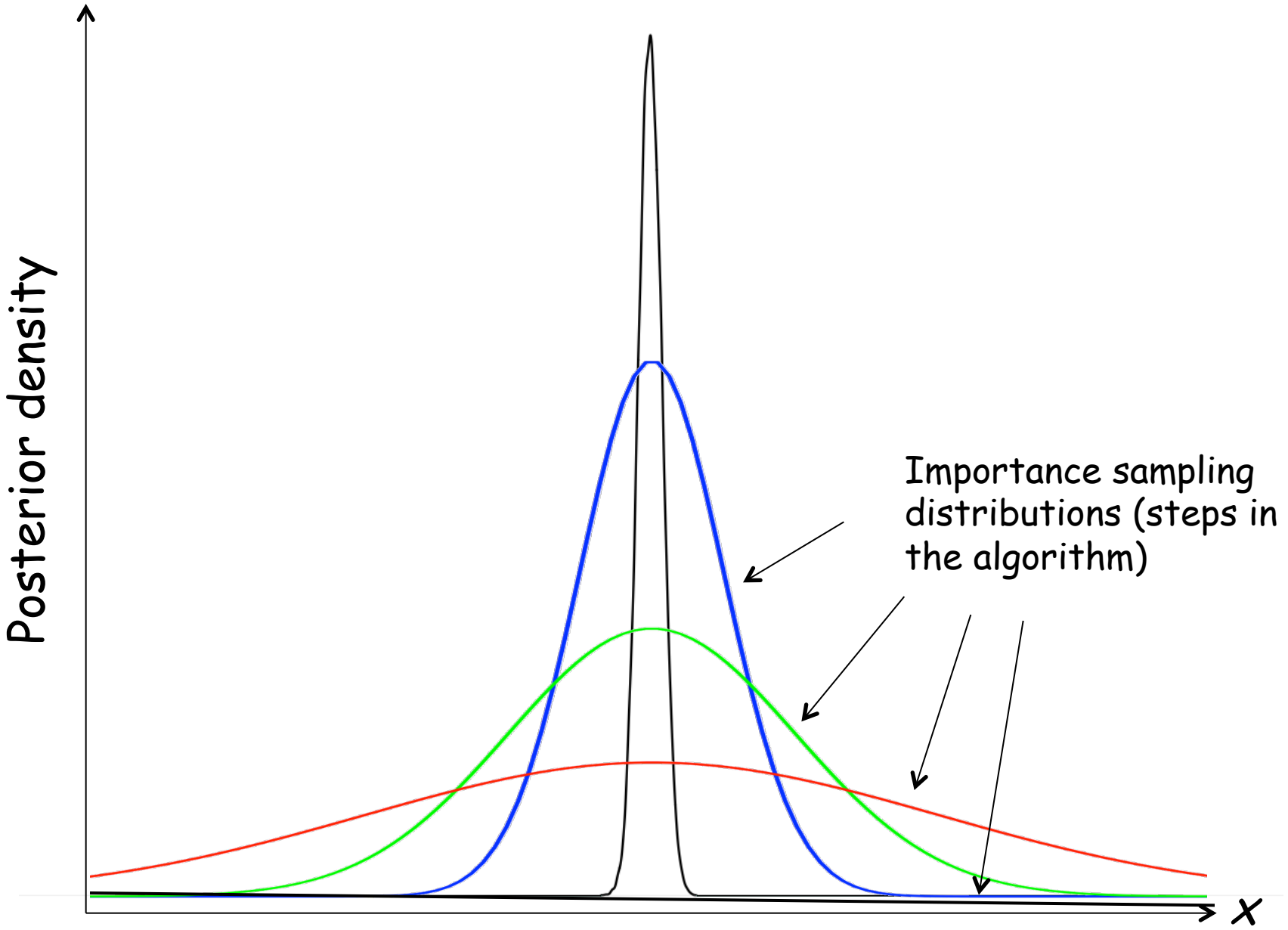
$$r_{ss} = \frac{c_{1.0}}{c_{0.8}} \cdot \frac{c_{0.8}}{c_{0.6}} \cdot \frac{c_{0.6}}{c_{0.4}} \cdot \frac{c_{0.4}}{c_{0.2}} \cdot \frac{c_{0.2}}{c_{0.0}}$$

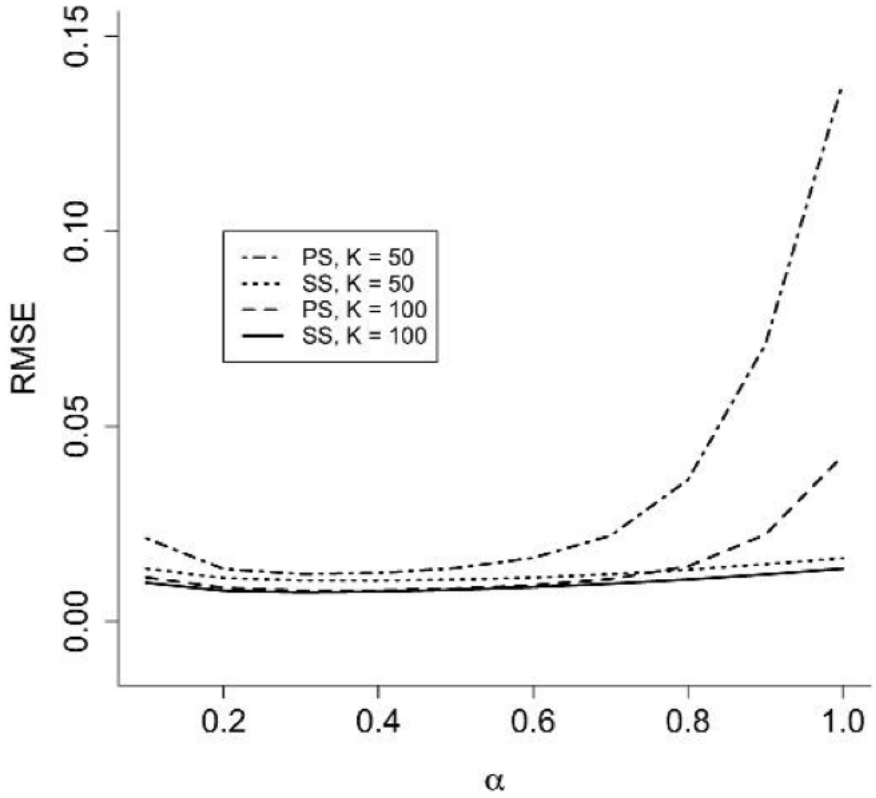
Stepping stone sampling bridging posterior and prior

Xie et al. 2011 Systematic Biology


Posterior density
X

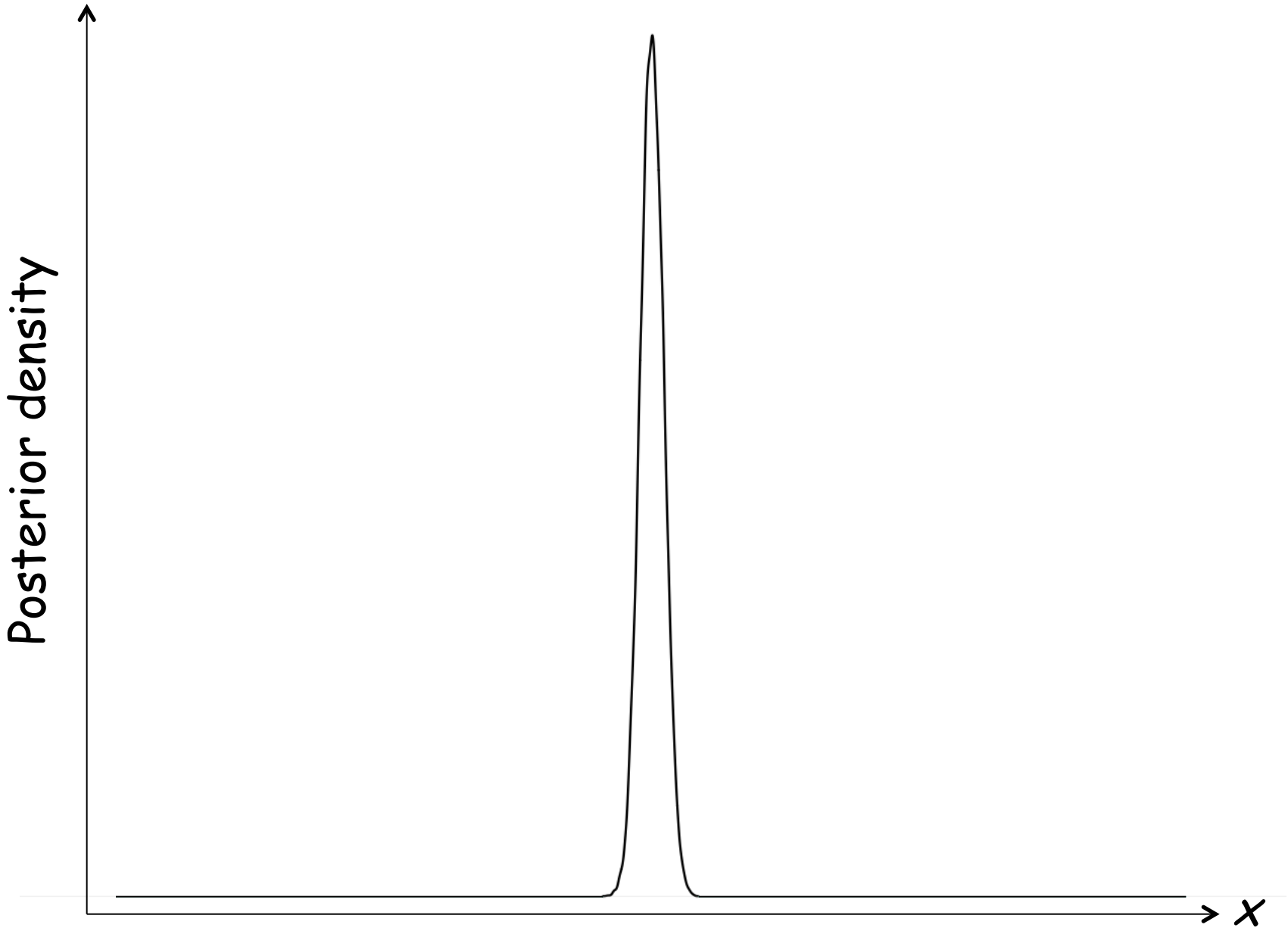

Posterior density
Importance sampling distributions (steps in the algorithm)
X

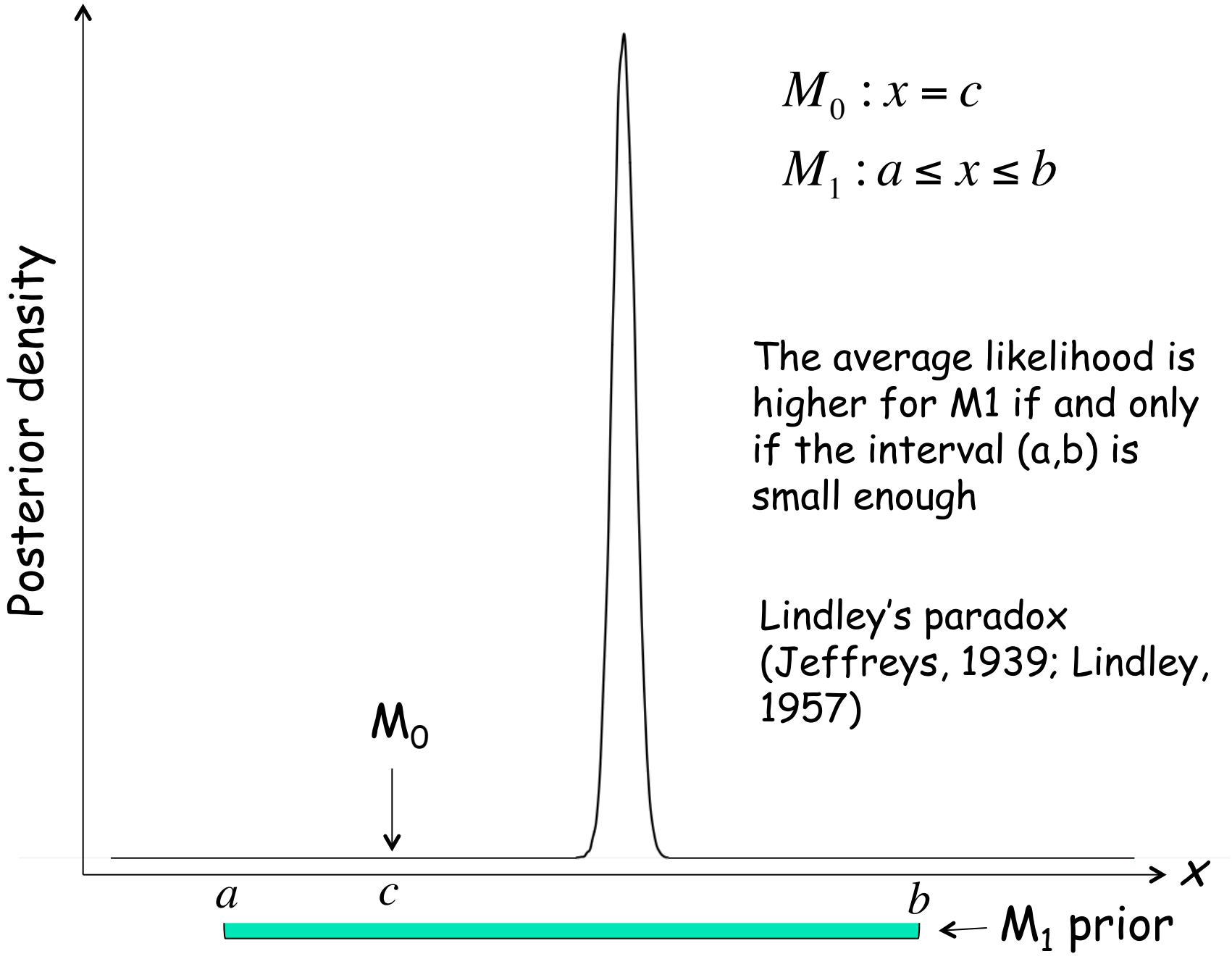

Posterior density
Importance sampling distributions (steps in the algorithm)
X



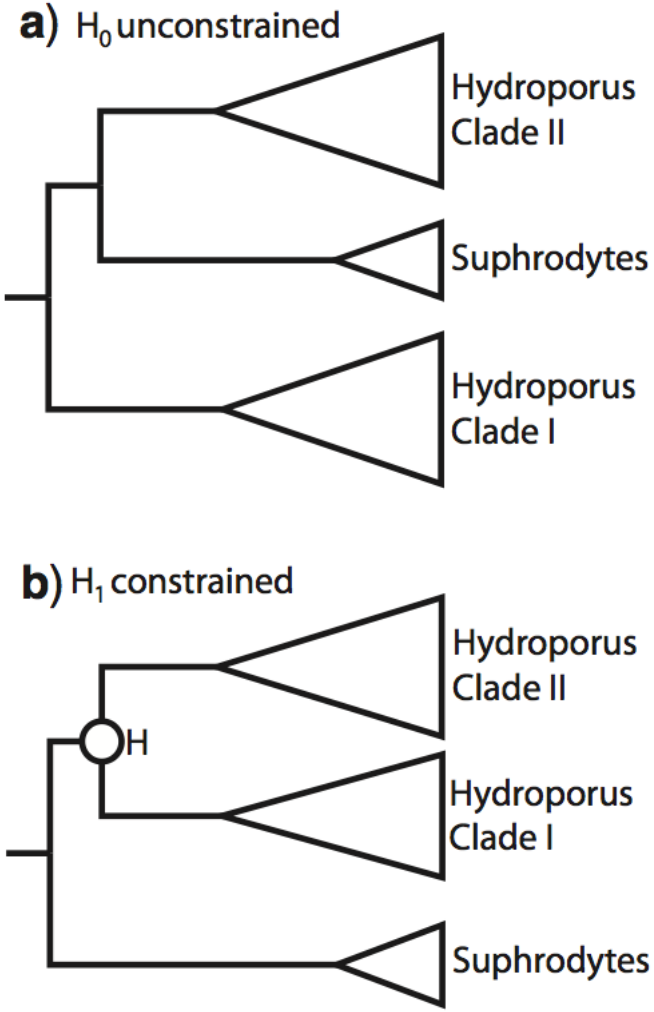
Performance of stepping stone sampling (SS) and thermodynamic integration (PS) for different numbers of steps and different stepping schemes (α). Smaller α values mean more samples close to the prior.


BAD NEWS!


Posterior density
X


$M_0 : x = c$
$M_1 : a \le x \le b$
The average likelihood is higher for M1 if and only if the interval (a,b) is small enough
Lindley's paradox (Jeffreys, 1939; Lindley, 1957)
Posterior density
M_0
a
c
b
x
M_1 prior

Bayesian tests of monophyly hypotheses



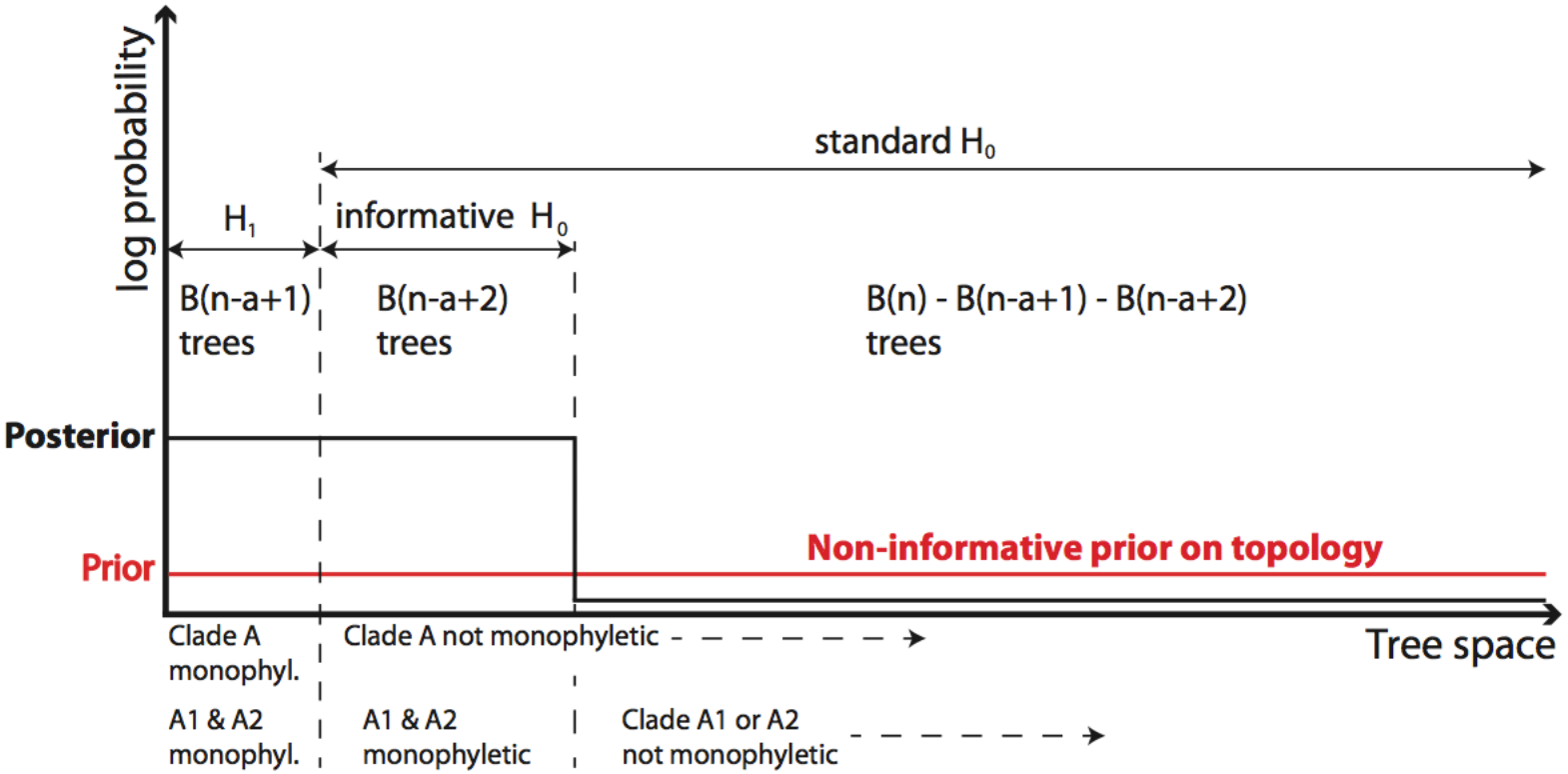
Posterior says there is strong support for *Hydroporus* **not** being monophyletic

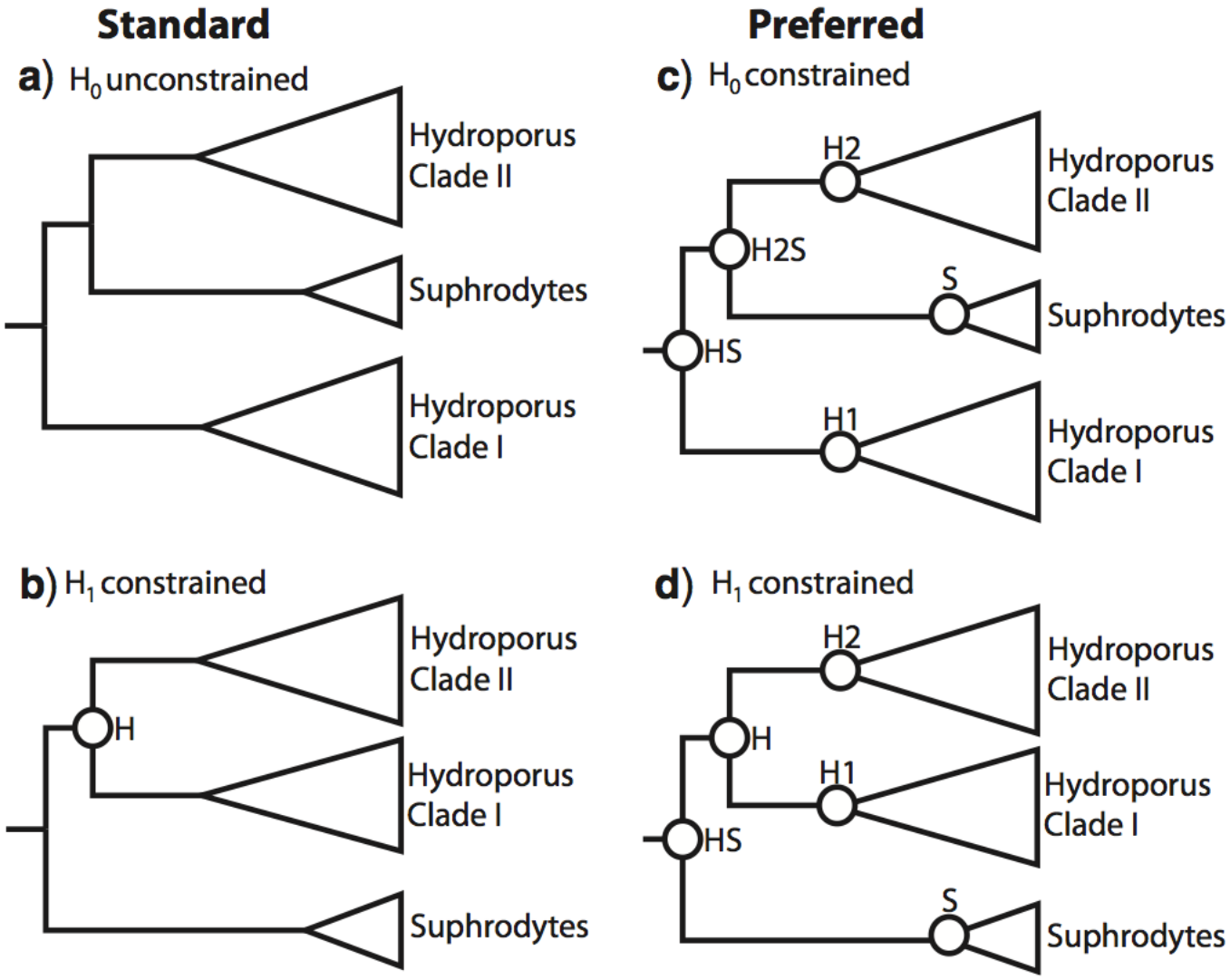
Bayes Factor test says there is very strong support for *Hydroporus* being monophyletic

What is going on?

Bergsten et al. 2013 Syst. Biol.

Schematic representation of posterior and prior





By constraining surrounding clades to be monophyletic, we can focus the prior topology space so that the Bayes factor test gives the expected result

2. The Model as a Random Variable

Conditioning on a model:

$$f(\theta \mid D, M) = \frac{f(\theta \mid M) f(D \mid \theta, M)}{f(D \mid M)}$$

Making the model a random variable:

$$f(\theta, M \mid D) = \frac{f(\theta, M) f(D \mid \theta, M)}{f(D)}$$

Model as random variable

- Advantage:
  - Only a single MCMC analysis
  - Computationally much less complex
  - Can integrate over much larger model spaces
- Disadvantages:
  - If models differ in dimensions, we need to use reversible jump MCMC algorithms; the dimension-matching requires some math
  - Even if dimensions are the same, it may be difficult to find proposals with good convergence (cf. topology inference problem)
  - Models with a "model-switching" parameter can have phase-transition behavior, complicating estimation of posterior distributions
- Bottom line: Wins in the end when you can do it

Substitution Model Space

GTR model parameters

$\boldsymbol{\pi}$	state frequencies	$\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$
$\boldsymbol{r}$	exchangeability rates	$\boldsymbol{r} = \{r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT}\}$

Model	Rate vector	Restr. Growth Fxn	K
GTR	$\boldsymbol{r} = \{r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT}\}$	{1,2,3,4,5,6}	6
HKY	$\boldsymbol{r} = \{r_{tv}, r_{ti}, r_{tv}, r_{tv}, r_{ti}, r_{tv}\}$	{1,2,1,1,2,1}	2
F81	$\boldsymbol{r} = \{r, r, r, r, r, r\}$	{1,1,1,1,1,1}	1

Huelsenbeck, Larget, Alfaro 2004 Mol Biol Evol

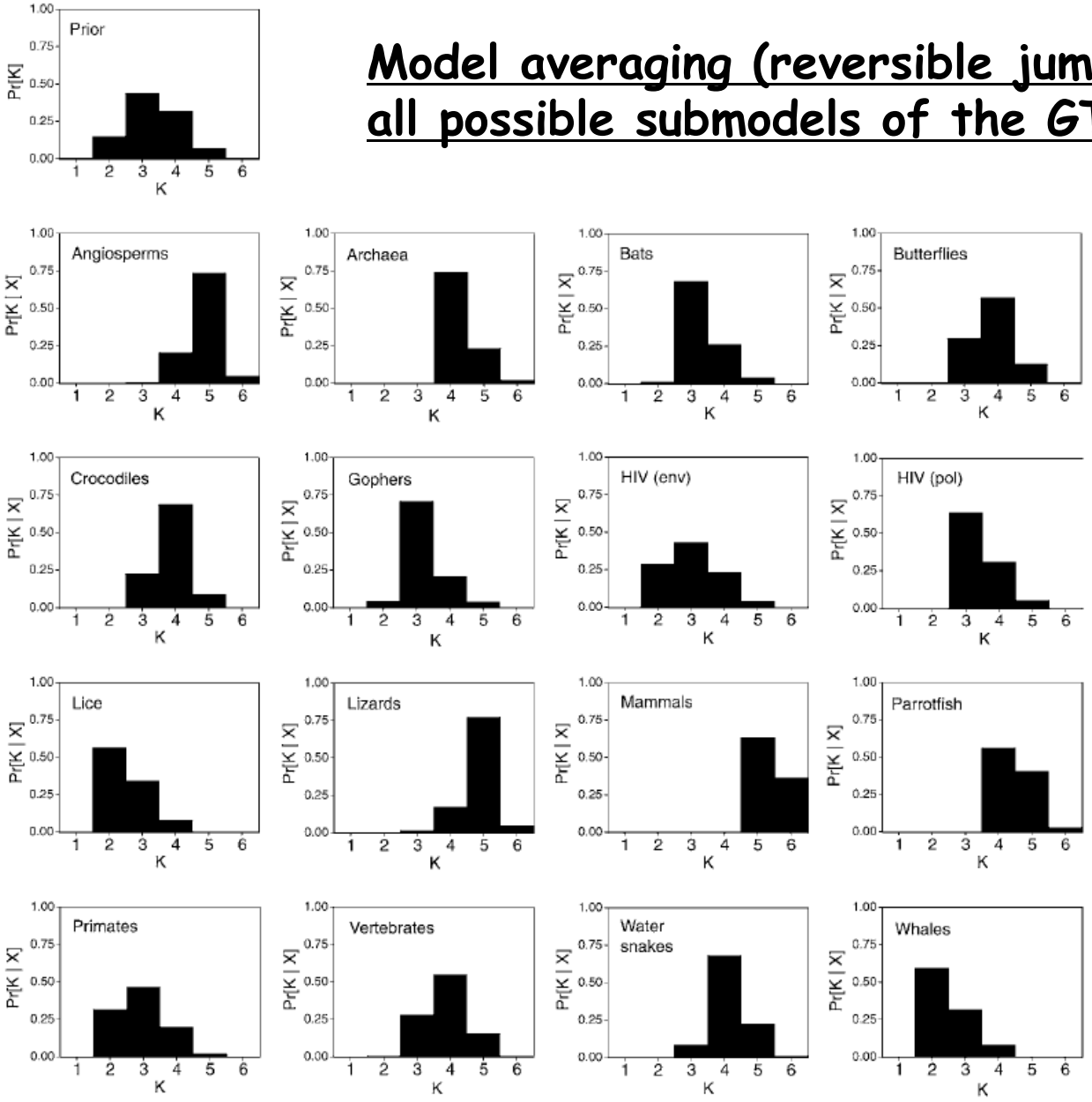
Table 1
All Possible Time-Reversible Models of DNA Substitution

K	Models				
1	$M_1 = 111111$				
2	$M_2 = 122222$	$M_3 = 121111$	$M_4 = 112111$	$M_5 = 111211$	$M_6 = 111121$
	$M_7 = 111112$	$M_8 = 112222$	$M_9 = 121222$	$M_{10} = 122122$	$M_{11} = 122212$
	$M_{12} = 122221$	$M_{13} = 122111$	$M_{14} = 121211$	$M_{15} = 121121$	$M_{16} = 121112$
	$M_{17} = 112211$	$M_{18} = 112121$	$M_{19} = 112112$	$M_{20} = 111221$	$M_{21} = 111212$
	$M_{22} = 111122$	$M_{23} = 111222$	$M_{24} = 112122$	$M_{25} = 112212$	$M_{26} = 112221$
	$M_{27} = 121122$	$M_{28} = 121212$	$M_{29} = 121221$	$M_{30} = 122112$	$M_{31} = 122121$
	$M_{32} = 122211$				
3	$M_{33} = 123333$	$M_{34} = 123222$	$M_{35} = 122322$	$M_{36} = 122232$	$M_{37} = 122223$
	$M_{38} = 123111$	$M_{39} = 121311$	$M_{40} = 121131$	$M_{41} = 121113$	$M_{42} = 112311$
	$M_{43} = 112131$	$M_{44} = 112113$	$M_{45} = 111231$	$M_{46} = 111213$	$M_{47} = 111123$
	$M_{48} = 122333$	$M_{49} = 123233$	$M_{50} = 123323$	$M_{51} = 123332$	$M_{52} = 123322$
	$M_{53} = 123232$	$M_{54} = 123223$	$M_{55} = 122332$	$M_{56} = 122323$	$M_{57} = 122233$
	$M_{58} = 121333$	$M_{59} = 123133$	$M_{60} = 123313$	$M_{61} = 123331$	$M_{62} = 112333$
	$M_{63} = 112322$	$M_{64} = 112232$	$M_{65} = 112223$	$M_{66} = 123122$	$M_{67} = 123212$
	$M_{68} = 123221$	$M_{69} = 121322$	$M_{70} = 121232$	$M_{71} = 121223$	$M_{72} = 122312$
	$M_{73} = 122321$	$M_{74} = 122132$	$M_{75} = 122123$	$M_{76} = 122231$	$M_{77} = 122213$
	$M_{78} = 123311$	$M_{79} = 123131$	$M_{80} = 123113$	$M_{81} = 121331$	$M_{82} = 121313$
	$M_{83} = 121133$	$M_{84} = 123211$	$M_{85} = 123121$	$M_{86} = 123112$	$M_{87} = 122311$
	$M_{88} = 122131$	$M_{89} = 122113$	$M_{90} = 121321$	$M_{91} = 121312$	$M_{92} = 121231$
	$M_{93} = 121213$	$M_{94} = 121132$	$M_{95} = 121123$	$M_{96} = 112331$	$M_{97} = 112313$
	$M_{98} = 112133$	$M_{99} = 112321$	$M_{100} = 112312$	$M_{101} = 112231$	$M_{102} = 112213$
	$M_{103} = 112132$	$M_{104} = 112123$	$M_{105} = 111233$	$M_{106} = 111232$	$M_{107} = 111223$
	$M_{108} = 112233$	$M_{109} = 112323$	$M_{110} = 112332$	$M_{111} = 121233$	$M_{112} = 121323$
	$M_{113} = 121332$	$M_{114} = 122133$	$M_{115} = 122313$	$M_{116} = 122331$	$M_{117} = 123123$
	$M_{118} = 123132$	$M_{119} = 123213$	$M_{120} = 123231$	$M_{121} = 123312$	$M_{122} = 123321$
4	$M_{123} = 123444$	$M_{124} = 123433$	$M_{125} = 123343$	$M_{126} = 123334$	$M_{127} = 123422$
	$M_{128} = 123242$	$M_{129} = 123224$	$M_{130} = 122342$	$M_{131} = 122324$	$M_{132} = 122234$
	$M_{133} = 123411$	$M_{134} = 123141$	$M_{135} = 123114$	$M_{136} = 121341$	$M_{137} = 121314$
	$M_{138} = 121134$	$M_{139} = 112341$	$M_{140} = 112314$	$M_{141} = 112134$	$M_{142} = 111234$
	$M_{143} = 123344$	$M_{144} = 123434$	$M_{145} = 123443$	$M_{146} = 123244$	$M_{147} = 123424$
	$M_{148} = 123442$	$M_{149} = 122344$	$M_{150} = 122343$	$M_{151} = 122334$	$M_{152} = 123423$
	$M_{153} = 123432$	$M_{154} = 123243$	$M_{155} = 123234$	$M_{156} = 123342$	$M_{157} = 123324$
	$M_{158} = 123144$	$M_{159} = 123414$	$M_{160} = 123441$	$M_{161} = 121344$	$M_{162} = 121343$
	$M_{163} = 121334$	$M_{164} = 123413$	$M_{165} = 123431$	$M_{166} = 123143$	$M_{167} = 123134$
	$M_{168} = 123341$	$M_{169} = 123314$	$M_{170} = 112344$	$M_{171} = 112343$	$M_{172} = 112334$
	$M_{173} = 112342$	$M_{174} = 112324$	$M_{175} = 112234$	$M_{176} = 123412$	$M_{177} = 123421$
	$M_{178} = 123142$	$M_{179} = 123124$	$M_{180} = 123241$	$M_{181} = 123214$	$M_{182} = 121342$
	$M_{183} = 121324$	$M_{184} = 121234$	$M_{185} = 122341$	$M_{186} = 122314$	$M_{187} = 122134$
5	$M_{188} = 123455$	$M_{189} = 123454$	$M_{190} = 123445$	$M_{191} = 123453$	$M_{192} = 123435$
	$M_{193} = 123345$	$M_{194} = 123452$	$M_{195} = 123425$	$M_{196} = 123245$	$M_{197} = 122345$
	$M_{198} = 123451$	$M_{199} = 123415$	$M_{200} = 123145$	$M_{201} = 121345$	$M_{202} = 112345$
6	$M_{203} = 123456$				

NOTE.—K is the number of substitution types. The named models are M_1, M_{15}, M_{40}, M_{122}, M_{168}, M_{195}, and M_{203}.

All 203 submodels of GTR

Model averaging (reversible jump MCMC) over all possible submodels of the GTR model



Posterior probability

Number of substitution types (K)

Huelsenbeck et al., 2004, MBE

Fixed number of substitution types

lset nst=1
lset nst=2
lset nst=6

Integrate over 203 models

lset nst=mixed

sump output

```
Model probabilities saved to file "replicase.nex.mstat".
Overwriting file "replicase.nex.mstat"

                          Posterior      Standard         Min.         Max.
Model                    Probability     Deviation     Probability    Probability
-------------------------------------------------------------------------------
gtrsubmodel[112234]         0.051          0.001         0.050         0.051
gtrsubmodel[121123]         0.335          0.013         0.326         0.344
gtrsubmodel[121134]         0.145          0.022         0.130         0.161
gtrsubmodel[123324]         0.063          0.008         0.057         0.069
-------------------------------------------------------------------------------
```

3. Model Adequacy

Model Adequacy

- Ideally, a model adequacy test would tell you whether a model is adequate for a particular set of observations
- Is there a way to tell whether a model is adequate?
- Answer is no. The best we can do is to test whether a model is inadequate from some particular perspective

Posterior predictive distribution

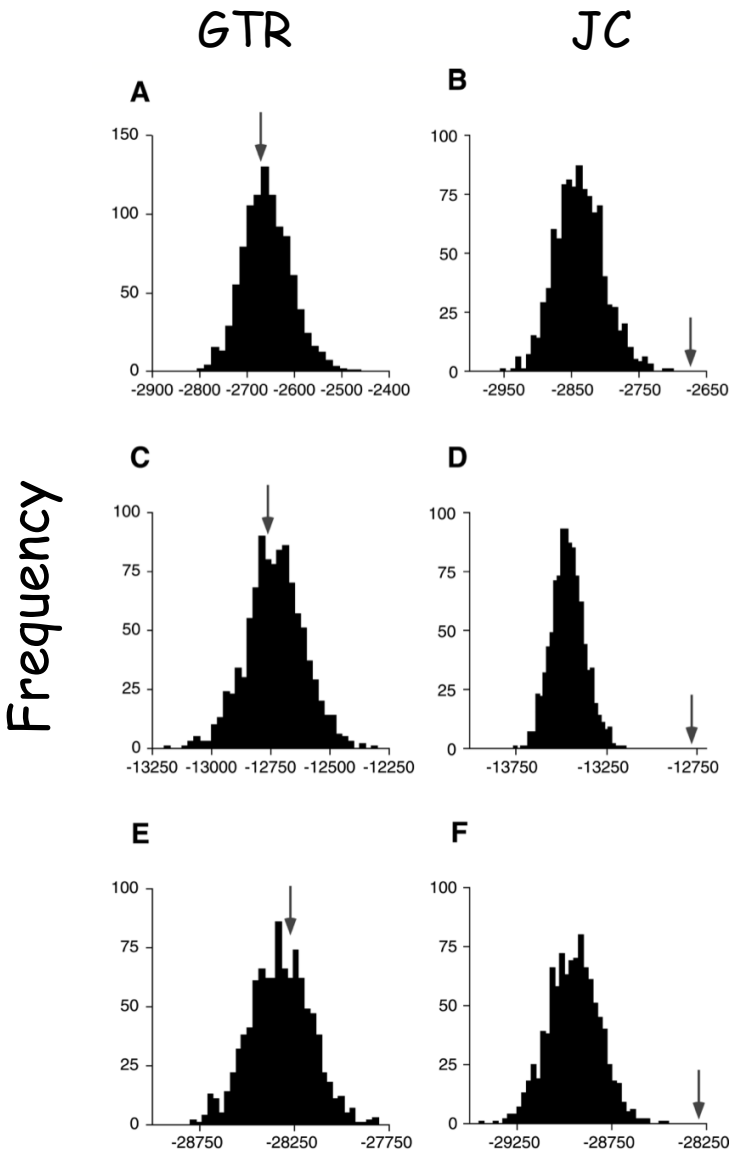
The predicted distribution of a new data point $\tilde{d}$

$$f(\tilde{d} \mid D) = \int f(\tilde{d} \mid \theta) f(\theta \mid D) \mathrm{d}\theta$$

That is, we can generate new data by pulling samples from the posterior, and then simulate one data point for each sample, using the parameter values for that sample.

Similar to parametric bootstrapping except that is more robust because it incorporates uncertainty in parameter estimates.

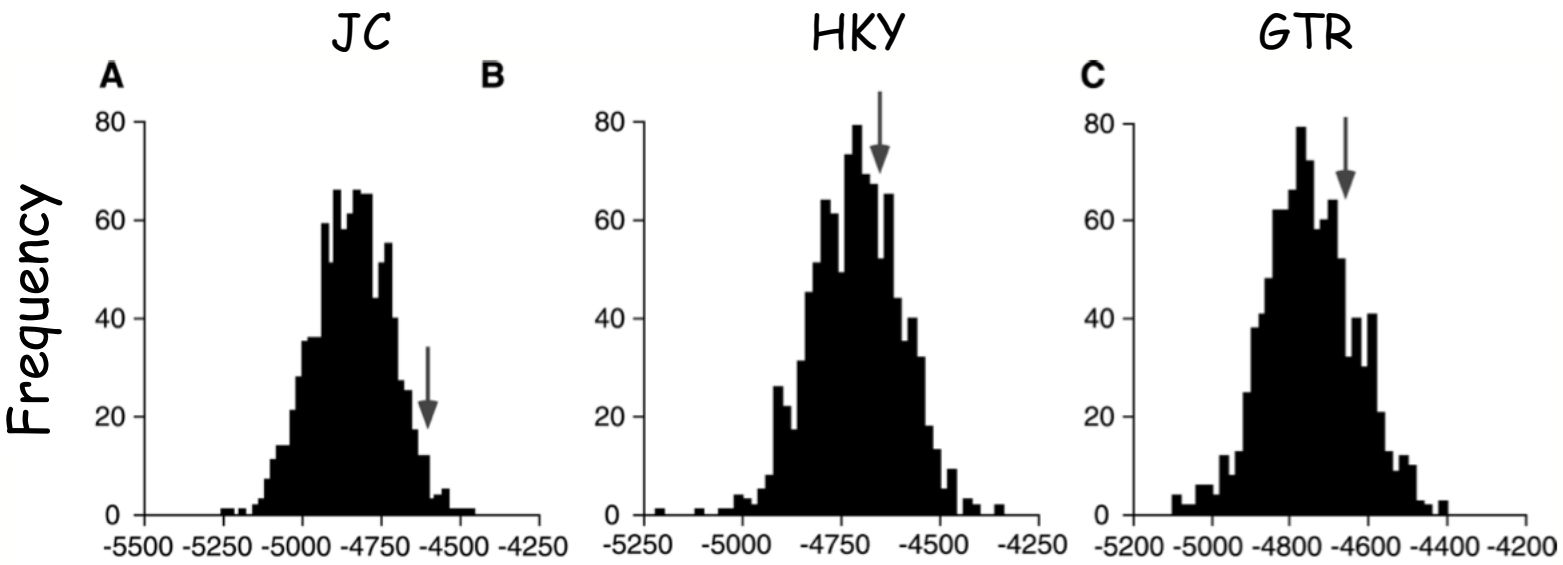
A posterior predictive test of model adequacy is based on generating new data sets from the posterior predictive distribution, and then comparing the value of a test statistic between the new data sets and the original data set.



Simulated data

Bollback, 2002, Mol Biol Evol

Empirical data



NB! No rate variation in these models

Bollback, 2002, Mol Biol Evol

Summary

- Bayes factors provide a powerful framework for comparing models:
  - Models need not be nested
  - No correction for number of parameters
  - Simpler models can win over more complex models if they have better average likelihoods
- Caveats:
  - Model likelihoods are difficult to estimate. Harmonic mean estimator is biased and more accurate methods, like stepping-stone sampling and thermodynamic integration, are time consuming
  - Lindley's paradox: Outcome is influenced by model priors -> beware!
- Treating the model as a random variable is a smarter approach, but it is technically more demanding
- Posterior predictive distribution can be used for model adequacy tests. Only inadequacy can be shown.
- Use a pragmatic approach as complement (or replacement):
  - Does the model make sense?
  - Is it possible to estimate the posterior generated by the model?
  - Does the model pick up signal in the data? That is, are posterior distributions different from priors? Does the signal make sense?